Can hybridization be detected between African wolf and sympatric canids?

Sunniva Helene Bahlk

Master of Science Thesis

2015

CEES

Centre for Ecological and Evolutionary Synthesis

Center for Ecological and Evolutionary Synthesis
Department of Bioscience
Faculty of Mathematics and Natural Science

University of Oslo, Norway

2015

Can hybridization be detected between African wolf and sympatric canids?

Sunniva Helene Bahlk

http://www.duo.uio.no/

Print: Reprosentralen, University of Oslo

Table of contents

Acknowledgments

This thesis was written at the Center for Ecological and Evolutionary Synthesis (CEES) at the Department of Biology, University of Oslo, under the supervision of Nils Christian Stenseth and Eli Knispel Rueness.

First I would like to thank my supervisors who trusted me with this project. Nils Christian, thank you for letting me be a part of CEES and for helping me get the permissions and privileges I needed to conduct this project. I am very grateful to be a part of this. And Eli, what would I have done without you? You gave me this amazingly interesting field of study, you guided me when I was blind, believed in me the times I lost hope, and pushed me when I needed it. Thank you for letting me spend countless hours in your office, ready to answer all possible questions, and for putting me in contact with all the right people that made it possible to realize my hopes and dreams for this project. Thank you for being an incredible supervisor.

I want to thank Robin Cristofari, who introduced me to several of the programs that became a huge part of this thesis. I am grateful for the assistance with ANGSD and all the other tools you helped me with. I admire you for your endless patience with all my questions, and for always giving me the feeling of being welcome. I want to thank Michael Matschiner for your ideas, knowledge and experience with bioinformatics, and for making my thesis better by running analyses on you unpublished program. I want to thank Emiliano Trucchi for joining the brainstorming at the beginning of the project.

I want to thank all my good friends, both at UiO and outside, for all the laughs, the outburst of frustration, the equally important meaningful and meaningless conversations, and the inspiring tea meetings. Thank you for helping me with my thesis, in big or small ways – it

means a lot to me. Thank you for your time, your motivational notes, the pep talks, and for believing in me. You know who you are.

I want to thank my family; you also believed in me and encouraged me every time I needed it, and even more. Thank you for always supporting me and for being so genuinely interested in what I am doing.

And last, but not least, I want to thank my loving Bård. Thank you for always being there for me, while simultaneously letting me use all my time on this project, leaving you as a single parent. Thank you for taking so good care of Brage, and for always trying to make my days even better. You are truly the best!

Abstract

Hybridization is a common phenomenon within the genus *Canis*. The recently discovered African wolf (*Canis lupaster*), is sympatric with several closely related species; the Ethiopian wolf (*Canis simensis*), the side-striped jackal (*Canis adustus*), and semi-domestic dogs (*Canis familiaris*). The aim of my thesis is to apply genome-wide data to investigate whether signs of hybridization can be detected between the African wolf and its sympatric canids. I used RAD-sequence data for 35 samples from Africa and 10 samples of grey wolf (*Canis lupus*) from North America. After demultiplexing and filtering the data from each sample, I kept 28 individuals for further analysis: ten African wolves, seven dogs, four Ethiopian wolves and seven grey wolves. I used the ANGSD (Analyzing Next Generation Sequencing Data) software for variant calling. This program is particularly suited to low or medium depth data as it takes genotype uncertainty into account. Various approaches were applied to study admixture and phylogenetic relationships among the species, and I was able to, for the first time, confirm the occurrence of hybridization between African wolf and dog, and between African wolf and Ethiopian wolf.

Introduction

The species in the genus *Canis* are closely related and widely distributed

The species in the genus *Canis* (wolf-like canids) are carnivore mammals found on all continents except Antarctica. One of the most commonly known members of this genus is the Holarctic grey wolf (*Canis lupus lupus*). They are found in the wilderness in the northern hemisphere and share habitats with canids like coyote (*Canis latrans*) and dog (*Canis lupus familiaris*). A close relative to the grey wolf is the newly discovered African wolf (*Canis lupaster*), which is found in northern, western and eastern parts of Africa. A mitochondrial DNA study showed that a formerly known golden jackal subspecies (*Canis aureus lupaster*) was in fact an unknown species more closely related to the grey wolf (*1*). Eurasian and African golden jackals are now confirmed to be distinct species through analyses of both nuclear and mitochondrial DNA, with the grey wolf as the African wolf's closest relative (*2, 3*). Compared to other canids, the African wolf has a high level of genetic and phenotypic diversity (*4*), and in some cases it can be difficult or impossible to distinguish an African wolf from the sympatric side-striped jackal (*Canis adustus*). In addition to the side-striped jackal, the African wolf is also sympatric with semi-domestic village dogs (*Canis lupus familiaris*) and the world's rarest canid, the Ethiopian wolf (*Canis simensis*). Compared to the large geographical range of the African wolf, the Ethiopian wolf is endemic to the highlands of Ethiopia. According to the IUCN's Red List, the Ethiopian wolf population is declining, with only 197 mature individuals left in 2013 (*5*).

These species are only some of the wolf-like canids. There are some uncertainties regarding the relationship between the species, their origin, and if some of the species rather should be classified as a subspecies (*6*). It still remains to describe a consensus phylogeny, and this could partly be because gene flow is confirmed among species in this genus (*7-10*).

Next-generation sequencing and bioinformatics

When looking for gene flow, the aim is to find species-unique alleles from one species represented in individuals of another closely related species. This can be done by comparing DNA samples and looking for variations and similarities between individuals and populations. The goal is to remove all genetic characters that are similar in both species, and just compare alleles that are unique to each species. If one individual contains some species-unique alleles from another species, it could be due to gene flow. To be able to perform studies like this, it is necessary to look at as much of the genome as possible. Genomics is a branch of biotechnology for genetic mapping and DNA sequencing of sets of genes or complete genomes of a selected organism. A range of new techniques makes it possible to conduct genomic studies at a reasonable cost. These modern techniques are defined by the umbrella term “Next-Generation Sequencing” (NGS).

NGS technologies are making a huge impact on many areas of biology, and have proven to be very suitable for detecting signs of hybridization (*11*). The term NGS is used to describe a number of modern sequencing tools, which are cheaper and faster than the previously used Sanger Sequencing (*12*). NGS can produce millions of small fragments on a single run, covering larger parts of the genome than before. NGS is also quicker compared to Sanger sequencing, and the accuracy is higher, which results in severely lowering the cost. Several different NGS technologies are available today; one of them is called Illumina (*13*). This method uses clonal amplification and sequencing by synthesis (SBS) chemistry. The process simultaneously identifies DNA bases while incorporating them into a nucleic acid chain. Each base emits a unique fluorescent signal as it is added to the growing strand, which is used to determine the order of their DNA sequence. To run Illumina sequencing, the input samples must be cleaved into short sections, typically fragments of 200-500 base pairs (BP). One method to collect these fragments is called Restriction-Site Associated DNA Sequencing (RADSeq) (*14*). RADSeq generates data sets of relatively short sequences from a large number of loci across the whole genome, from several individuals at the same time. DNA from each individual is cut with a chosen restriction enzyme, producing a set of sticky-end fragments. An adapter (P1), containing an Illumina adapter and a molecular identifier (MID, also called “barcode”), is attached to the cut site. Samples from multiple individuals are pooled together, and the tagged fragments are randomly sheared. The result of the shearing

is that only a subset of the resulting fragments contains restriction site and P1 adapter. Each fragment is selected by size, and the sheared fragments of approximately 200-500 bases are ligated to a second adapter (P2), which has a divergent "Y" structure. The next step is to run a Polymerase Chain Reaction (PCR) to amplify the sheared and marked fragments. The PCR uses two primers that bind to each adapter. The "Y" structure of the P2 adapter ensures that the PCR amplification will only happen if the fragment is completed with a P1 adapter. The result is that the amplified DNA contains an Illumina adapter, MID, the partial restriction site, a few hundred bases of flanking sequence and a P2 adapter. This RADSeq library will be sequenced on the Illumina platform. The sequence is generated from the MID in the P1 adapter and across the restriction site generating a data set of RAD-tags from the whole genome. Each sequence is called one read. RADSeq can generate millions of reads, and the likelihood of collecting many Single-Nucleotide Polymorphisms (SNPs) is high. After sequencing, this huge amount of data depends on high performance computational resources to be processed. Because of ever-developing technologies, this is now possible.

Bioinformatics was a term introduced in 1970 to distinguish the information processing area of biology from other areas such as biophysics and biochemistry (*15*). In the late 80's the term was changed to describe "*computational methods for data management and data analysis*" (*15*). This includes a combination of computer science, statistics, mathematics and engineering in order to analyze and interpret biological data. As well as being an umbrella term for the biological studies that use computer programming as a part of the methodology, bioinformatics may also refer to specific pipelines that are repeatedly used in fields like genomics and genetics. Such pipelines often include optimizing the readability of the data, collecting SNPs, comparing samples, analyzing them, and doing statistical tests. Today there are hundreds of different kinds of software and tools available for different use and new and more complex methods will constantly be made available due to ongoing technological development. While some of the programs offer a graphical user interface (GUI), which makes it easier to exploit this development in the field of biology, a majority of the programs only offer a command line interface (CLI). The user must write commands to the program in the form of successive lines of text (command lines). Although there are several hundred tools available, bioinformaticians often need to write their own scripts because the existing tools are not sufficient for their particular use. This can be challenging

for a user that lacks knowledge and experience with informatics, due to poorly documented tools and with few standards to follow. In order to write the correct commands and use the right settings, it is necessary to understand how the tools work. It is also important to know enough about the biology behind the case to be able to choose the right tool in the first place.

The concepts of hybridization and introgression

Hybridization can be defined as interbreeding of individuals from taxonomically different populations (*16*). When gene flow continues through backcrossing of hybridized individuals to one or both of their parental populations, and there is a stable integration of genetic material from one species into another, it is defined as introgression (*17*). In some cases it can be hard to identify hybrids, particularly in the case of introgression. Phenotypically, a hybridized individual can be identical to one of the parental species (*18*), and at gene level introgression and incomplete lineage sorting can seem confusingly similar (*19*). Incomplete lineage sorting is a phenomenon where a polymorphic ancestral species divides into two lineages where some of the same alleles are retained in the descendant branches (*20*). In the cases where the gene tree differs from the population tree, the explanation could be both incomplete lineage sorting and introgression. To detect introgression, and to distinguish it from incomplete lineage sorting, it is necessary to do statistical tests.

Some degree of gene flow may occur between closely related species. Several occasions of hybridization between species in the *Canis* genus has been reported. Hybridization events between grey wolf and feral dogs have been genetically verified in Italy (*9, 21*), Estonia (*22*), Latvia (*22*), Scandinavia (*23*), Georgia (*24*), Spain (*25*) and Canada (*26*). Grey wolf and coyote hybrids are common in North America (*7*), and a litter by a female Ethiopian wolf and a male feral dog has been reported in Ethiopia (*8*). Mating events between Himalayan wolf (*Canis himalayensis*) and feral dog is reported in India (*27*), even though no offspring have been observed yet. Hybridization and introgression between Red wolf (*Canis rufus*) and coyote is considered as the biggest threat for the conservation of the endangered Red wolf (*28*). Both in case studies like these, and generally speaking, it has been suggested that hybridization can play an important role in evolution (*29*).

The aim of my study

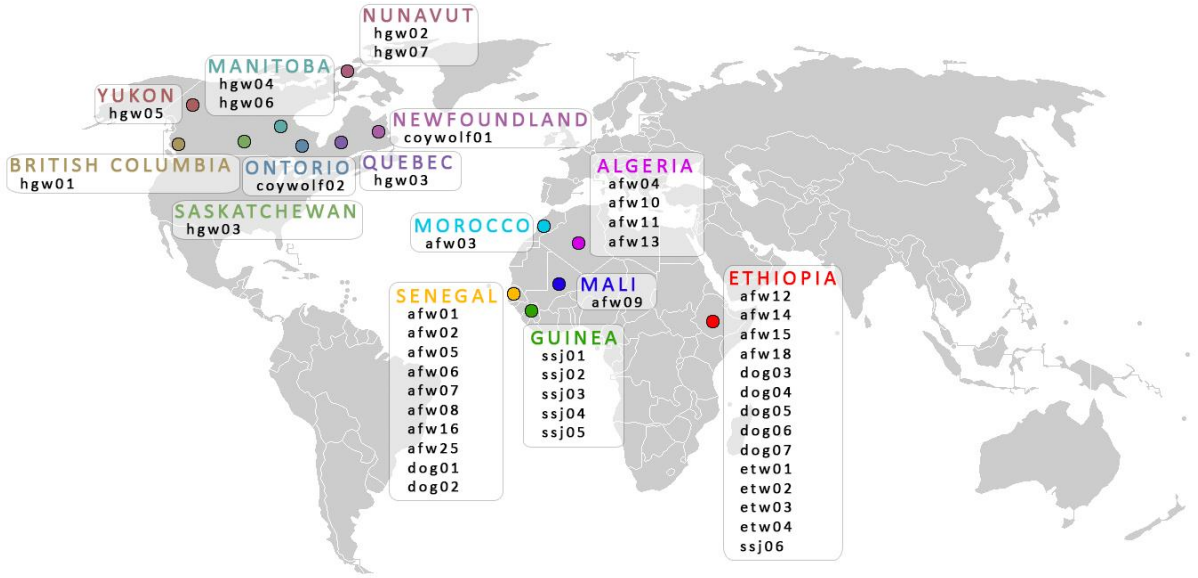
Using a range of different bioinformatics tools, the aim of my study was to investigate whether hybridization and introgression occurs between the African wolf and its sympatric canids. I planned to separate individual samples from two RADSeq libraries, align them to the most relevant reference genome, and clean each sample from content that could be misleading in order to achieve optimized readability. I would take genotype uncertainty into account by calculating likelihoods for each allele, collecting SNPs, and analyzing these results from several different angles.

Materials and Methods

Origin of the samples and laboratory protocols

My supervisor, Eli Rueness, received 45 tissue, blood and pelt samples of canids from collaborators in France[1], England[2], and Denmark[3]. They have contributed samples from North America (N=10), Ethiopia (N=14), Senegal (N=10), Guinea (N=5), Algeria (N=4), Morocco (N=1), and Mali (N=1). The samples were from 18 African wolves ("afw"), eight Holarctic grey wolves ("hgw"), seven dogs ("dog"), six side-striped jackals ("ssj"), four Ethiopian wolves ("etw"), and two "coywolves" (coyote/grey wolf hybrids).
(See figure 1 and table 1 for complete overview.)

Figure 1 – A map illustrating the locations where the samples were collected.



[1] Philippe Gaubert, Institut des Sciences de l'Evolution de Montpellier (ISEM), France.
[2] Claudio Sillero-Zubiri, University of Oxford, England.
[3] Mikkel Sinding, Statens Naturhistoriske Museum, Denmark.

DNA was extracted in the CEES-lab by Eli Rueness using the Qiagen DNeasy Blood & Tissue Kit. The DNA concentration of each sample was quantified by a fluorometric-based method (Qubit 2.0, Life technologies) and diluted to 0.5µM (see table 1). Restriction Site Associated DNA Sequencing (RADSeq) was used to generate millions of fragments following a protocol based on the publication from Baird et al. (*30*). The samples were divided into two libraries, RAD2 and RAD3. RAD3 included 13 re-sequenced samples of individuals that were already represented in RAD2. The restriction enzyme *SbfI* (NEB) was used to cut the DNA into fragments. A set of 22 uniquely barcoded P1 adapters (MID) was used in the first library (RAD2) (table 1). In the second library (RAD3), two sets of barcoded adapters (MID) were used: 12 P1 adapters and three P2 adapters. A defined number of cycles with sonication were used to shear the libraries. The target size of the fragments was 300–500 BP and size selection was performed using a BluePippin instrument (Sage Science). PCR (Polymerase Chain Reaction) was used to amplify the DNA in each library. The DNA concentrations of the amplified libraries were quantified once more by Qubit, and an Agilent Bioanalyzer chip (Invitrogen) was also used to check the molarity. Finally, a volume of 20 µl per library (with a DNA concentration of 45 ng/µl) was sequenced on the Illumina HiSeq2000 platform of the Norwegian Sequencing Center (University of Oslo). The sequence length was 120 BP for both forward and reversed reads.

Format of the resulting files

The sequenced libraries were downloaded in a fastq-format; with two files per library (one for each reading direction). Fastq is a text format for storing nucleotide sequences and their corresponding quality score. The file is a long list of four lines per read. The first line starts with the "@" character followed by an identifier or optional description. In my case all samples were marked with their unique barcode. The second line is the raw sequenced letters. The third line contains a "+" character and an optional description. Line four encodes the quality values for the sequence in line two, where each symbol indicates the quality of the individually sequenced base. The files were uploaded, stored and processed at the ABEL computing cluster.

Table 1 – The table is a complete list of all 58 samples included in the two libraries. The table is separated in two: the first part contains the samples from the RAD2 library, and the second part contains the samples represented in RAD3. All samples are ordered by their label.

Column 1: The *label* of each sample.
Column 2: The *library* the sample comes from.
Column 3: The *species* the sample is phenotypically identified as.
Column 4: The *location* where the sample was collected.
Column 5: The barcode used in the *P1 adapter*
Column 6: The barcode used in the *P2 adapters* from the second library (RAD3).
Column 7: The amount of DNA, *µg/ml DNA quantification,* in each sample.

Label	Library	Species	Location	P1 adapter	P2 adapter	DNA q. µg/ml	
afw01	RAD2	*Canis lupaster*	Senegal	AATTT		36,9	[1]
afw02	RAD2	*Canis lupaster*	Senegal	ACACG		42,3	[1]
afw03	RAD2	*Canis lupaster*	Morocco	ACCAT		19,0	[1]
afw04	RAD2	*Canis lupaster*	Algeria	AGTCA		127,0	[1]
afw05	RAD2	*Canis lupaster*	Senegal	ATCGA		13,8	[1]
afw06	RAD2	*Canis lupaster*	Senegal	ATGCT		202,0	[1]
afw07	RAD2	*Canis lupaster*	Senegal	ATTAG		86,4	[1]
afw08	RAD2	*Canis lupaster*	Senegal	CAACT		9,4	[1]
afw09	RAD2	*Canis lupaster*	Mali	CATGA		43,8	[1]
afw10	RAD2	*Canis lupaster*	Algeria	CCAAC		22,2	[1]
afw11	RAD2	*Canis lupaster*	Algeria	CCCCA		10,3	[1]
afw12	RAD2	*Canis lupaster*	Ethiopia	CCGGT		14,7	[2]
afw13	RAD2	*Canis lupaster*	Algeria	CGCGC		10,5	[1]
afw25	RAD2	*Canis lupaster*	Senegal	CAGTC		17,3	[1]
dog01	RAD2	*Canis familiaris*	Senegal	AAAAA		375,0	[1]
dog02	RAD2	*Canis familiaris*	Senegal	AACCC		54,3	[1]
ssj01	RAD2	*Canis adustus*	Guinea	AAGGG		23,5	[1]
ssj02	RAD2	*Canis adustus*	Guinea	ACGTA		94,1	[1]
ssj03	RAD2	*Canis adustus*	Guinea	ACTGC		30,0	[1]
ssj04	RAD2	*Canis adustus*	Guinea	AGAGT		63,2	[1]
ssj05	RAD2	*Canis adustus*	Guinea	AGCTG		73,2	[1]
ssj06	RAD2	*Canis adustus*	Ethiopia	CGATA		58,7	[2]

[1] Philippe Gaubert, Institut des Sciences de l'Evolution de Montpellier (ISEM), France.
[2] Claudio Sillero-Zubiri, University of Oxford, England.

Table 1 - continued

Label	Library	Species	Location	P1 adapter	P2 adapter	DNA q. µg/ml	
afw02-2	RAD3	*Canis lupaster*	Senegal	CACGGT	GACT	42.0	[1]
afw03-2	RAD3	*Canis lupaster*	Morocco	CGTTAG	CTGAT	59.4	[1]
afw04-2	RAD3	*Canis lupaster*	Senegal	ACCTGA	GACT	127.0	[1]
afw06-2	RAD3	*Canis lupaster*	Senegal	CGTTAG	GACT	202.0	[1]
afw07-2	RAD3	*Canis lupaster*	Senegal	CAGTCT	GACT	86.0	[1]
afw09-2	RAD3	*Canis lupaster*	Mali	GATGCG	GACT	43.8	[1]
afw10-2	RAD3	*Canis lupaster*	Algeria	TTACTC	GACT	22.0	[1]
afw12-2	RAD3	*Canis lupaster*	Ethiopia	GCACTA	GACT	14.7	[2]
afw13-2	RAD3	*Canis lupaster*	Algeria	ATGGAC	GACT	11,0	[1]
afw14	RAD3	*Canis lupaster*	Ethiopia	TGCACT	ACTT	75.8	[2]
afw15	RAD3	*Canis lupaster*	Ethiopia	ACCTGA	ACTT	54.0	[2]
afw16	RAD3	*Canis lupaster*	Senegal	ACCTGA	CTGAT	31.7	[1]
afw18	RAD3	*Canis lupaster*	Ethiopia	TTACTC	CTGAT	35.0	[2]
afw25-2	RAD3	*Canis lupaster*	Senegal	GTATCG	GACT	17.0	[1]
coywolf01	RAD3	«Coywolf»	Newfoundland, CA	CACGGT	CTGAT	76.9	[3]
coywolf02	RAD3	«Coywolf»	Ontario, CA	TGCACT	CTGAT	59.4	[3]
dog01-2	RAD3	*Canis lupaster*	Senegal	AGTCAC	GACT	375.0	[1]
dog02-2	RAD3	*Canis lupaster*	Senegal	TCGATA	GACT	54.0	[1]
dog03	RAD3	*Canis familiaris*	Ethiopia	AGTCAC	ACTT	31.7	[2]
dog04	RAD3	*Canis familiaris*	Ethiopia	TCGATA	ACTT	65.6	[2]
dog05	RAD3	*Canis familiaris*	Ethiopia	GTATCG	ACTT	68.8	[2]
dog06	RAD3	*Canis familiaris*	Ethiopia	GATGCG	ACTT	52.2	[2]
dog07	RAD3	*Canis familiaris*	Ethiopia	CGTTAG	ACTT	170.0	[2]
etw01	RAD3	*Canis simensis*	Ethiopia	GCACTA	CTGAT	7.7	[2]
etw02	RAD3	*Canis simensis*	Ethiopia	CAGTCT	CTGAT	7.0	[2]
etw03	RAD3	*Canis simensis*	Ethiopia	CACGGT	ACTT	410.0	[2]
etw04	RAD3	*Canis simensis*	Ethiopia	ATGGAC	ACTT	84.6	[2]
hgw01	RAD3	*Canis lupus*	British Columbia, CA	TTACTC	ACTT	238.0	[3]
hgw02	RAD3	*Canis lupus*	Nunavut, CA	GCACTA	ACTT	23.0	[3]
hgw03	RAD3	*Canis lupus*	Quebec, CA	CAGTCT	ACTT	570.0	[3]
hgw04	RAD3	*Canis lupus*	Manitoba, CA	AGTCAC	CTGAT	254.0	[3]
hgw05	RAD3	*Canis lupus*	Yukon, CA	TCGATA	CTGAT	64.9	[3]
hgw06	RAD3	*Canis lupus*	Manitoba, CA	GTATCG	CTGAT	374.0	[3]
hgw07	RAD3	*Canis lupus*	Nunavut, CA	GATGCG	CTGAT	252.0	[3]
hgw08	RAD3	*Canis lupus*	Saskatchewan, CA	ATGGAC	CTGAT	1000.0	[3]
ssj02-2	RAD3	*Canis adustus*	Guinea	TGCACT	GACT	94.0	[1]

[1] Philippe Gaubert, Institut des Sciences de l'Evolution de Montpellier (ISEM), France.

[2] Claudio Sillero-Zubiri, University of Oxford, England.

[3] Mikkel Sinding, Statens Naturhistoriske Museum, Denmark.

The process of filtering the libraries and the individual files

Separating libraries into individual files:

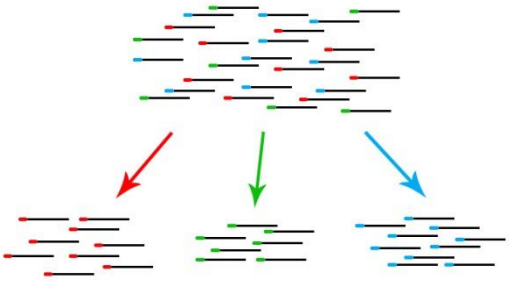


Figure 2 – An illustration of how each RAD-tag is recognized and separated by a sample-unique barcode.

Since all samples were stored together in four files, the pipeline started with separating each sample from the libraries. This is called demultiplexing (figure 2). The raw data were processed on the RAD1 server using a software pipeline called Stacks v1.28 (*31*). Within Stacks, there is a program called process_radtags, which examines raw reads from Illumina sequencing runs. The program was executed twice, once for each library. Since the libraries contained two fastq-files each, process_radtags had two different sets of input files. The first input file in both runs was the fastq-file containing the reads sequenced from the forward DNA strand (-1 input_forward.fastq.gz). The second input file contained the reads from the reverse strand (-2 input_reversed.fastq.gz). Process_radtags runs through the data, and selects reads with a high enough quality that other programs will be able to process them. The first criterion is that the sample-unique barcode and RAD-cut site are intact (-c, for “clean”). If they are intact, process_radtags will continue the demultiplexing procedure. If an error occurs in the barcode or the RAD site, process_radtags can correct them (-r, for “rescue”) within a certain allowance. Next, the program will slide a window down the length of the read and check the average quality score within the window. This window covers 15% of the length of the read. If the quality drops below 90% probability of being correct, the read will be discarded (-q, for “quality score”). If the read is accepted, it will be stacked in a file containing only reads with the same barcode. In this way Stacks will make an individual file for each recognized barcode. Because of the more advanced use of barcodes in RAD3, I included a setting in Stacks to give the new generated files a predetermined name instead of the barcode (--inline_inline). The identifier for each read (previously using the barcodes), was changed to a more complex identifier including reading direction. Since both forward and reverse DNA strands were sequenced, it was generated one new fastq file per strand for every individual.

```
~/stacks-1.28/process_radtags \
-i gzfastq \                          #input format
-1 input_forward.fastq.gz \           #forward DNA input
-2 inpur_reversed.fastq.gz \          #reversed DNA input
-y fastq \                            #output format
-o output \                           #output destination and prefix
-b barcodes.txt \                     #list of barcodes in each sample
-e sbfI \                             #restriction enzyme used
--inline_inline \                     #use labels instead of barcodes
-c \                                  #clean reads
-r \                                  #rescue errors
-q                                    #discard bad quality reads
```

After demultiplexing, I removed eight individuals that were identified (by phenotype) as either side-striped jackal or coyote/grey wolf hybrids ("coywolf"). The side-striped jackal is not relevant in this study because this species is too distantly related to the focal species. The "coywolf" was removed since I wanted to focus on hybridization events in Africa, not North America, and because I did not have any coyotes to compare them with. Such hybrids could give misleading results.

Alignment to reference genome:

The next step for the remaining individuals was to compare them to each other. When comparing genomes, it is necessary to know where each read originated. This is done by alignment. Flicek and Birney (*32*) defined alignment as "*the process of determining the most likely source within the genome sequence for the observed DNA sequencing read, given the knowledge of which species the sequence has come from. Sequencing reads may also be aligned to other genomes, assuming the evolutionary distance between the genomes is appropriate*". To achieve this I chose the program Bowtie 2 v2.2.4 (*33*) and the most recently released dog genome (CanFam 3.1 (*34*)) as the reference (setting –x in Bowtie 2). Bowtie 2 is a fast and memory-efficient aligning tool, which output the files in SAM (Sequence Alignment Map) format. This format is a generic format for sorting large nucleotide sequence alignment, simplifies the process of enabling a larger number of tools later on. The two fastq-files per sample generated by Stacks' process_radtags were used as input files in

Bowtie 2 (-1 for forward reads and -2 for reverse reads). I used the --fr setting which is appropriate for Illumina's Paired-end Sequencing Assay. --fr indicates that an alignment is valid only if there is a candidate paired-end alignment where mate 1 appears upstream of the reverse complement of mate 2 and the fragment length constraints are met. I set the minimum fragment length for valid pair-end alignments to be 250 (-I) and the maximum fragment length to 1000 (-X). I disabled discordant alignments (--no-discordant) since I only allowed paired end alignment (--fr). The definition of discordant alignment is an alignment where both mates align uniquely which is impossible when combined with --fr. I also disabled Bowtie 2's function of finding individual mates when no concordant or discordant alignment is found (--no-mixed). Since the Illumina pipeline uses ASCII characters, I chose the equivalent "Phred+33" encoding (--phred33). The last thing I defined in Bowtie 2 was to suppress SAM records for reads that failed to align (--no-unal). The goal with these settings was to remove all aligned reads without a proper paired-end candidate, and alignments that could be aligned to more than one region. Without removing those reads, I would lose the opportunities that many bioinformatics tools provide. In addition to this, it can be impossible to compare individuals if there is no information about the origin of the reads. All these settings were included in the alignment of every individual. To avoid having to manually run Bowtie 2 on each sample, I wrote a looping script that collected each sample name from a text-file, and generated an array. For each item in the array (i.e. each sample name) the script would automatically pick up a new set of input files, and run through the defined settings in Bowtie 2, and generate an output file based on the inputs. This looping resulted in one converted and aligned SAM file (-s) for each pair of fastq files.

```
readarray array < ~/individuals_list.txt
arr="${array[*]}"

for e in $arr; do
      if [[ "$e" == *"afw"* ]] ; then s="afw"; fi
      if [[ "$e" == *"dog"* ]] ; then s="dog"; fi
      if [[ "$e" == *"etw"* ]] ; then s="etw"; fi
      if [[ "$e" == *"hgw"* ]] ; then s="hgw"; fi

      echo -e "\n\n\tProcessing sample ""$e""...";

      ~/bowtie2-2.2.4/bowtie2 \
            -x reference_genome \
            --fr -I 250 -X 1000 --no-discordant \
            --no-mixed --phred33 --no-unal \
            -1 ~/data/*/fastq/$e.1.fq \
            -2 ~/data/*/fastq/$e.2.fq \
            -S ~/data/SAM/$e.sam

      echo -e "\tDone";
      echo "-----------------------------------";

done
echo "Finish!"
```

Merging of samples of the same individual represented in both libraries:

Since some of the individuals were represented in both libraries, the same individuals were now also represented in two different SAM files. Each of these samples had been through the laboratory protocol twice, and the probability of sequencing different reads is relatively high. Merging the two files produced a single file that contains both input records. I used a function called “merge” in SAMtools v1.1 (*35, 36*) to do this. SAMtools is a library and software packages used to manipulate alignments in the SAM/BAM format. BAM is the binary version of SAM. A binary file is not a text file like SAM, but a computer file. It is readable, but not human readable. Even though the BAM file can contains the same information as the SAM file, the computer format will sort the numeric data in a way that saves space and computational effort. SAMtools is designed to manipulate BAM and SAM files equally, but most other programs prefer one of the formats.

Filtering and cleaning each sample:

In order to optimize the readability of each SAM file, they had to be cleaned before starting comparison and analysis. Three different programs were used in this process; SAMtools v1.1, Picard v1.113 (*37*), and R v3.2.1 (*38*). All the filtering was processed through a script ("RAD_BAM.sh") consisting of eight parts, written by Robin Cristofari.

1) The first part of the filtration pipeline was to set a minimum for mapping quality. Mapping quality can be defined as uniqueness and indicates the probability that the selected read is aligned correctly. For instance, a read that originated inside a repeat element might align equally to numerous regions in the genome, leaving the aligner with no basis for preferring one over the other. I chose 30 as the minimum mapping quality, which means that there is a 1 in 30 chance that the read truly originated elsewhere. All reads with a higher chance of being misplaced was discarded. This were done with Samtools-1.1 function view –q.

```
#Filtering with minimum MAPQ
$SAMTOOLS/samtools view -h -q $MAPQ \
      -S $INPUT_DIR/$SAMPLE.sam \
      -o $OUTPUT_DIR/BAM/$SAMPLE.mapq.sam
```

2) The second part of the filtration was done in R. After removing some of the reads in step 1, paired reads may become orphaned. Orphaned reads can make the overall quality lower and some programs cannot read such files. To avoid this, I removed all reads that were not properly paired. The file name was changes (--trim and mv) and the no longer needed input files was deleted (rm).

```
#Filter out orphaned reads
$SCRIPTS/SAM_KeepOnlyPairs.R \
      --S=$OUTPUT_DIR/BAM/$SAMPLE.mapq.sam \
      --out=$OUTPUT_DIR/BAM/ --trim --embed \
      2>&1 | tee $OUTPUT_DIR/log/$SAMPLE.orphans.log

mv $OUTPUT_DIR/BAM/$SAMPLE.mapq.sam.pairs $OUTPUT_DIR/BAM/$SAMPLE.pairs.sam
rm $OUTPUT_DIR/BAM/$SAMPLE.mapq.sam
```

3) The now fully paired samples were converted from SAM-format to BAM with SAMtools. This did not change any content, but was necessary prior to the next step in the filtration process.

```
#Converting SAM to BAM
$SAMTOOLS/samtools view -b
      -S $OUTPUT_DIR/BAM/$SAMPLE.pairs.sam
      -o $OUTPUT_DIR/BAM/$SAMPLE.pairs.bam

rm $OUTPUT_DIR/BAM/$SAMPLE.pairs.sam
```

4) Picard v1.113 is a set of Java tools for working with next generation sequencing data in BAM format. Since the SAM/BAM format allows storing several individuals in the same file, there is a slot where it is possible to define which read group the sequenced read originates from. Since the files in this case only stored one individual each, and SAMtools will not allow an empty spot, the slots were filled with some default information. This was done with a function called AddOrReplaceReadGroups in Picard.

```
#Adding read groups
java -jar $PICARD/AddOrReplaceReadGroups.jar \
      I= $OUTPUT_DIR/BAM/$SAMPLE.pairs.bam \
      O= $OUTPUT_DIR/BAM/$SAMPLE.group.bam \
      LB= RAD-SAMPLE \
      PL= ILLUMINA \
      PU= RADSEQ \
      SM= $SAMPLE \
      QUIET=TRUE \
      VERBOSITY=ERROR

rm $OUTPUT_DIR/BAM/$SAMPLE.pairs.bam
```

5) The fifth step was going back to SAMtools and sorting the reads in each BAM file by the leftmost coordinates, in this case the identifier of each read. The content of the coordinate was identical to the identifier in the individual fastq files produced by Stacks' process_radtags.

```
#Sorting reads
$SAMTOOLS/samtools sort
      $OUTPUT_DIR/BAM/$SAMPLE.group.bam
      $OUTPUT_DIR/BAM/$SAMPLE.sort

rm $OUTPUT_DIR/BAM/$SAMPLE.group.bam
```

6) The newly sorted BAM-files were taken back to Picard to remove PCR duplicates. Duplicates are created during the PCR amplification and when identical PCR products are sequenced multiple times. The main purpose of removing duplicates is to mitigate the effects of PCR amplification bias since PCR duplicates do not contain any new information. There is also a computational benefit to reducing the number of reads to be processed in the downstream steps. MarkDuplicates is a tool within Picard that detects and removes these duplicates.

```
#Removing duplicates
java -jar $PICARD/MarkDuplicates.jar \
      INPUT=$OUTPUT_DIR/BAM/$SAMPLE.sort.bam
      UTPUT=$OUTPUT_DIR/BAM/$SAMPLE.dedup.bam \
      METRICS_FILE=$OUTPUT_DIR/log/$SAMPLE.dedup.metrics \
      REMOVE_DUPLICATES=true \
      READ_NAME_REGEX='[0-9]_([0-9]+)_([0-9]+)_([0-9]+)_paired' \
      QUIET=TRUE \
      VERBOSITY=ERROR

rm $OUTPUT_DIR/BAM/$SAMPLE.sort.bam
```

7) In all prior steps, the selected reads I wanted to keep or discard were stored in temporary files. The content of these files were exported to two new BAM files: one with the accepted reads and one with the discarded reads.

```
#Export only the selected reads
$SAMTOOLS/samtools view \
      -f 1 \
      -b $OUTPUT_DIR/BAM/$SAMPLE.dedup.bam \
      -o $OUTPUT_DIR/BAM/$SAMPLE.$TYPE.bam
```

8) The last step was to put an index in the BAM files containing the reads I wanted to keep. To get quick random access to the data, the regional data were indexed to be able to limit the user interface of the SAMtools view function and similar commands to particular regions of interest.

```
#Index the final BAM file
$SAMTOOLS/samtools index $OUTPUT_DIR/BAM/$SAMPLE.$TYPE.bam
```

Since the cleaning process removed unwanted reads, the total number of reads was reduced in all files. Depending on the quality of the sample, the number was highly variable. To decide which samples were of good enough quality to be included, I checked the total number of reads left in each file. I used the view -c (-c for “count”) function of SAMtools and decided to exclude African wolves that had a total read number below 2.500.000 and grey wolves that had a total read number below 400.000. The remaining individual would then go on to the next step which was to extract polymorphic sites and compare the individuals to each other.

Analyzing the genotypes and visualizing the results

SeqMonk v0.30.2 (*39*) is a program for visualizing how the reads of each sample or individual maps to the reference genome. Available assembled reference genomes are implemented in the program, but the reference genome used in the aligning process (in Bowtie 2) must be the same genome as the one used in SeqMonk. Since the only available dog genome in SeqMonk was an earlier published version (CanFam 2.0), I contacted the producers (Babraham Bioinformatics). Just a few hours later, CanFam 3.1 was available in the program. The graphical user interface (GUI) consists of a data panel, genome overview and the chromosome viewer. The data panel contains a series of folders with the various annotations and the data sets I imported (BAM files). The data panel can be used to change different settings, generate plots, and perform analyses. The genome overview shows a graphical representation of the whole dog genome laid out in chromosomes. It is possible to click and drag anywhere and the region selected will be shown in the chromosome viewer. The chromosome viewer is the most detailed view of the chromosome and provides

information about each annotated gene from the reference genome. Underneath the feature tracks are a series of white and grey data tracks and each of these will contain the data for each BAM file imported. The information shown in these tracks can be changed; in my case they illustrated each read. I took some screenshots from the program to illustrate how different reads align to the dog genome. I did not use any quantitating or filtering tools that are included in the program since I had other applications I found more suitable for those tasks.

ANGSD v0.901 (Analyzing Next Generation Sequencing Data) (*40*) is the software I used to analyze the BAM files and generate output files that could be used in further analysis. The main reason why I chose ANGSD as the tool to collect the variable sites is because it takes genotype uncertainty into account. This is especially useful for low and medium depth data. Another reason why I chose ANGSD is that it includes implementations of large sets of downstream analyses that are not implemented in any other software. This includes allowing the user to choose from multiple methods for intermediate analysis, and ANGSDs ability to correspond to a variety of other programs. These features make it easier and faster to conduct desirable analysis. A final advantage of ANGSD the use of less wall clock time (not CPU) compared with similar NGS analysis programs (like SAMtools and GATK). ANGSD uses the information from the BAM files to either process the individuals alone or compare the individuals with each other. Based on the settings in ANGSD, different outputs were produced. These outputs are mostly enormous text files containing genotype likelihoods or called genotypes. I ran the program three times with different settings. Two general settings with their associated functions were included in all three runs: i) genotype likelihood (GL (*36*)) and ii) allele frequency (doMAF (*41*)). i) Genotype likelihood is an important part of ANGSD and is included in most of the programs possible functions. Genotype likelihood calculation is based on the aligned reads, associated mapping, and sequencing quality score. This likelihood is the marginal probability of the sequencing data given a genotype in a particular individual and in a particular site of the genome. Instead of just collecting the sampled alleles, it takes genotype uncertainty into account and calculates the likelihood that these are correct. Genotype uncertainties can arise from sources such as mapping and sequencing errors and the random sampling of haploid reads from a diploid genotype.
ii) Allele frequency calculation is the relative frequency of an allele for a site, and it can be

predetermined which of the alleles are major or minor. The dog genome (CanFam 3.1) was used as a reference for this. I could choose between using the reference genome as an ancestor or just as a reference. I chose ancestor because the reference function presupposes a folded genome spectrum. A folded genome spectrum contains only half the information of a non-folded spectrum, and I did not want to lose this potentially important information about allele frequencies. To reduce space and computation time, I ordered the program to only collect bases with a frequency with a p-value less than $1e^{-6}$. These bases contain the most information about the variation between the individuals and are therefore the most important focus.

The goal of the first run in ANGSD was to generate input files for a Principal Component Analysis (PCA). A PCA is a statistical procedure to analyze the variation between components in different dimensions. To do this, I had to estimate the site allele frequency likelihood (-doSaf 1 (*42*)). This calculation was based on the individual genotype likelihood assuming HWE (Hardy-Weinberg Equilibrium). I also did a genotype calling (-doGeno 32), which dumps the binary posterior for all samples. To do this I had to estimate the posterior genotype probability based on the allele frequency as a prior (-doPost 1).

```
# The first run in ANGSD
./angsd \
      -bam ~/bam.list \
      -GL 1 \
      -P 16 \
      -SNP_pval 1e-6 \
      -doGeno 32 \
      -doPost 1 \
      -doMaf 2 \
      -doSaf 1 \
      -anc ~/Canis_familiaris/genome.fa \
      -doMajorMinor 5 \
      -out ~/output_prefix
```

The output file with the extension .geno.gz was used as an input file in ngsCovar. ngsCovar is a tool within a tool package called ngsPopGen (Next-Generation Sequencing Population Genetics) (*43*). The program's function is to generate a covariance matrix between pairs of individuals. I included settings to remove non-variable and low-coverage sites (-call 0), filter

out sites with estimated map allele frequency (MAF) less than 0.05, and had the program run through every site in the input files (-nsites 2127714). The covariance matrix generated by ngsCovar was used as an input file in R. A script for producing cluster information and a plot were included in the ngsPopGen package. I specified that I was interested in the two first dimensions of the PCA plot (-c 1-2), since these are the two that explains the most variability between the individuals. I made some changes to the scripts in order to include the names of each sample in the plot. I also added a confidence ellipse defining 95% of the confidence interval within each species.

```
# Generating covariance file
      ./nobackup/ngsPopGen/ngsCovar \
      -probfile input.geno \
      -outfile output.covar \
      -nind 28 \
      -nsites 2127714 \
      -call 0 \
      -minmaf 0.05

### Generate a cluster file in R
individ<-scan("population_list.txt", what="", sep="\n")
write.table(cbind(
      seq(1,28), rep(1,28),
      c(rep("African wolf",10), rep("Dog",7), ("Ethiopian wolf",4),
            rep("Grey wolf",7)),
      c(individ)),
      row.names=F,
      sep="\t",
      col.names=c("FID","IID","CLUSTER","NAME"),
      file="cluster_name.clst",
      quote=F
)


### R-script for producing PCA plot
Rscript --vanilla --slave ./plotPCA.R -i input_covar.covar -c 1-2
      -a input_cluster.clst -o output_pca.pdf

### qqplot in the R-script
ggplot(data=PC, aes_string(x=x_axis, y=y_axis, color="Pop", label=Ind)) +
      geom_point() +
      geom_text(aes(label=Ind),hjust=1.2, vjust=0, size=3) +
      ggtitle(title) +
      stat_ellipse(data = PC, type="t", level = 0.95)
```

The second run in ANGSD was to generate an input file for an admixture analysis. The purpose of an admixture analysis is to quantify and visualize the admixture proportions in individuals based on a specified number of populations (clusters). This was a less complex run in ANGSD containing only the beagle generator (doGlf) in addition to the settings that were included in all ANGSD runs.

```
### The second run in ANGSD
./angsd \
      -bam ~/bam.list \
      -GL 1 \
      -nThreads 24 \
      -SNP_pval 1e-6 \
      -doGlf 2 \
      -doMaf 2 \
      -anc ~/Canis_familiaris/genome.fa \
      -doMajorMinor 5 \
      -out ~/output_prefix
```

The beagle.gz output is a genotype likelihood file that was used as an input file in a program called ngsAdmix (*44*). This program is an empirical-Bayes algorithm, and can easily be trapped on a local likelihood optimum. So to be more confident about the results, I did a bootstrap analysis. With an R-script (Bootstrap-Beagle.R, written by R. Cristofari), I generated 50 bootstrap replicates based on the output file made by ANGSD. Because of the bootstrap, the order of the clusters was randomly selected in all files. I made a loop that produced a matrix with admixture proportions per output file from the bootstrap. Each of these files is the input file for an admixture plot. Ideally I would have created a script that summed the results from every admixture proportion matrix, and automatically produced an accurate plot. The script should detect the correct, randomly ordered column based on the content of the consistent order of individuals. Each column should then get an ID and the script should restructure the content of each row based on the ID of the column. The goal is to have a consistent structure of the matrix. The next step in the script should be to summarize and calculate the support for the most common admixture proportions in each position of the matrix, and based on these numbers generate a new plot. The new plot would preferably contain the degree of support for each admixture proportion. Due to lack of time and scripting experience, I was not able to write a script that would perform such a

task. Instead I made an individual R-plot for each admixture matrix produced (supplementary figure 2a and 2b) to get a sense of any possible trend. I manually calculated the median value for each admixture proportion for every individual and made a new admixture plot based on those numbers. The script for the R-plot was delivered by ngsAdmix, but I made some graphical changes such as specifying the colors, defining a distance between each bar, and including labels on each sample.

```
### Bootstrap output file from ANGSD
./Bootstrap-Beagle.R \
      --likes=./input.beagle.gz \
      --bootstrap=50 \
      --out=./output

      ### Generate admixture matrix for each bootstrap with NGSadmix
for R in {1..50}
do
  ./NGSadmix \
  -likes input_file_$R.lhoods \
  -K 4 \
  -P 20 \
  -o admixture_output_$R \
  -minMaf 0.05
done

### Generate admixture plot
admix<-t(as.matrix(read.table("ngsAdmix_output.qopt")))
sample<-c(
      "afw02","afw04","afw06","afw07","afw09","afw10","afw13","afw16",
      "afw18","afw25","dog01","dog02","dog03","dog04","dog05","dog06",
      "dog07","etw01","etw02","etw03","etw04","hgw01","hgw03","hgw04",
      "hgw05","hgw06","hgw07","hgw08")

      barplot(
      admix,
      col=c('green','red','blue','orange'),
      border=NA,
      ylab="Admixture",
      names.arg=sample,
      main="Admixture plot",
      las=2,
      cex.names=1.6)
```

The last run in ANGSD was to make a Variant Call Format (VCF) file. This is a common text format in bioinformatics for sorting gene sequence variance. It is flexible and well adapted to large-scale genomics. In order to produce a VCF file, I needed both a numeric genotype

calling and frequencies. The genotype calling printed out major and minor alleles as "-1,0,1,2". These were calculated from the reference genome (-doGeno 3). I also needed posterior genotype probability based on the allele frequency as a prior (-doPost 1). The last task was to calculate the frequency of the different bases, A, T, C, and G (-doCounts 1). These frequencies were dumped in a separate file with chromosome number, position and depth (-dumpCounts 2). The depth is the sum of reads covering a site for all individuals.

```
### The third run in ANGSD
./angsd \
      -bam ~/bam.list \
      -GL 1 \
      -P 16 \
      -SNP_pval 1e-6 \
      -doPost 1 \
      -doGeno 3 \
      -doMaf 2 \
      -doCounts 1 \
      -dumpCounts 2 \
      -anc ~/Canis_familiaris/genome.fa \
      -doMajorMinor 5 \
      -out ~/output_prefix
```

The genotype calling and genotype frequency output files were used in an R-script (angsd2vcf.R, written by R. Cristofari) to convert the content to a VCF file. This VCF file was then converted to a distance matrix with a program called Plink v1.90 (*45*). Plink is a toolset designed for whole genome analysis with a focus on analysis of genotype and phenotype data. Since the standard setting is based on the human genome, I had to allow for extra chromosomes (--allow-extra-chr), and set the chromosome number to 38 (--chr-set 38). I defined the output format to be a distance matrix (--distance-matrix).

```
### Generating a distance matrix with Plink
./plink-1.90/plink
      --allow-extra-chr \
      --chr-set 38 \
      --distance-matrix \
      --vcf /vcf_file.vcf \
      --out output_prefix
```

The distance matrix generated by Plink was used in a nexus file. Nexus files always begin with the characters `#nexus`, but are otherwise organized into major units known as blocks. These blocks can among other include taxa blocks, data blocks, tree blocks or other so-called private blocks, which are only recognized by special programs. The distance matrix would be categorized as a data block and it is a recognizable format for SplitsTree4 (*46*). Based on this nexus file, I generated a neighbor-net to visualize the kinship between the individuals and possible splits. Neighbor-net is an algorithm for constructing phylogenetic networks that is based on Neighbor-Joining. Neighbor-net provides a snapshot of the data that can guide more detailed analysis (*47*). SplitsTree4 has a graphical user interface (GUI), and I downloaded a free trial to my computer. Because I was using the free trial version, I was not able to make changes in the plot. Instead, I manually changed the size of the text and added colored figures to indicate species and location in Photoshop CS6.

The VCF file was also converted to a special TreeMix format used by TreeMix v1.12 (*48*). A python script written by Michael Matschiner was responsible for the conversion. TreeMix is a program for inferring the patterns of population splits and mixtures in the history of a set of populations. The program will search for the maximum likelihood graph through two major steps. First, for a given topology, the program will try to find the maximum likelihood branch length and migration weights. Second, the program will search for new possible topologies through a hill-climbing approach. TreeMix will randomly choose three populations, optimize the branch length for all three possible trees and choose the best tree. In case of more than four populations, the program will continue to add populations one by one and search for the best possible tree. Since I only had four populations, where I specified Ethiopian wolf to be the outgroup, TreeMix did not try to add new populations. Instead the program continued with the first step of finding maximum likelihood branch length and migration weights. Before adding migration edges to a tree, it is important to set the position of the root (-root etw). I also had to define how many migration events I would allow (-m 1).

```
### Generate TreeMix plot
./treemix \
      -i ~/inputfile.txt.gz \
      -root etw \
      -m 1 \
      -o ~/output_prefix

### R
source("~/src/plotting_funcs.R")
plot_tree("input_prefix")
```

The same input file used in TreeMix was also used in a statistical program called F4 (*49*). F4 was used to test how much support there is for distinguishing introgression from incomplete lineage sorting. F4-statistics was introduced in 2009 by Reich, Thangaraj, Patterson, Price and Singh (*50*) and is a measurement to distinguish introgression from incomplete lineage sorting between four populations, A, B, C, and D. The population topology (collected from TreeMix) is assumed to be (A,B),(C,D); the F4-statistic is calculated as the product of the difference in allele frequencies between A and B, and between C and D. In case of incomplete lineage sorting, we would expect that the differences between A and B should be independent of those between C and D, and the F4 value should be zero. In the case of introgression, we would expect non-zero F4 values. As well as calculating the F4 value, the F4 program also calculates the support for the introgression. This process is called downweighting. The program runs simulations on Single Nucleotide Polymorphism (SNP) datasets with migration rates set to zero, and therefore strictly without introgression. The program would then run F4-statistics on each simulated dataset. If the value is more extreme than the first observed F4, the program would report the proportion. This difference would be a part of the measurement for how much support there is for the observed F4 value. When removing one individual at the time and then calculating a new F4-statistic, the program can find the individual that contributes the most to introgression on a population level. Since this individual has the highest F4 value, the total F4 value will drop. The output files contained the original F4 value and the downweighted F4 value. The program was run by M. Matschiner and I had the output files delivered.

Figure 3 – Graphical illustration of the methodic pipeline.

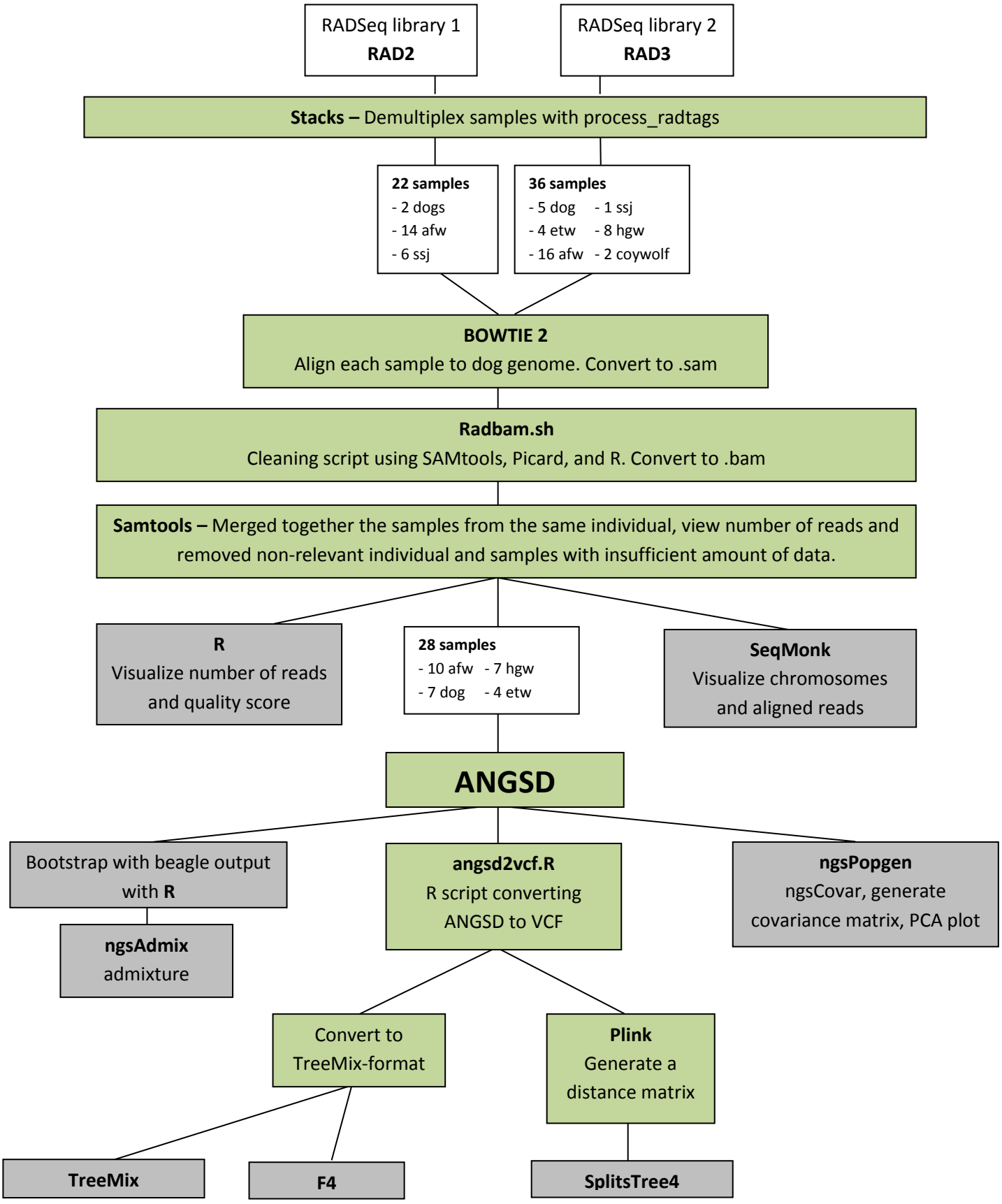


Table 2 - This table lists various software and tools used in the process. The name of the program and a short description is included.

Program name	Short description
ANGSD (*40*)	ANGSD is a software for analyzing next generation sequencing data. Most methods take genotype uncertainty into account instead of basing the analysis on called genotypes. But the program can do both. Collecting genotype likelihood is especially useful for low and medium depth data. The software is written in C++ and was used for genotype likelihood, genotype calling, allele frequencies and base frequencies.
Bowtie 2 (*33*)	Bowtie 2 is a tool for aligning sequenced reads to long reference sequences. Bowtie 2 supports gapped, local, and paired-end alignment modes. The program converts fastq files to SAM.
F4 (*49*)	F4 calculates the F4-statistic from allele frequencies of four populations and uses coalescent simulations to test whether this value could be the result of incomplete lineage sorting.
FastQC (*51*)	FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. The program was tested, but not used in the final pipeline.
ngsAdmix (*44*)	ngsAdmix is a tool for finding admixture proportions from NGS data. ngsAdmix is based on genotype likelihoods. It is a multithreaded c/c++ program from the same producers as ANGSD.
ngsPopGen ngsCovar (*43*)	ngsPopGen is a tool pack designed to analyze NGS data for population genetics purposes. ngsCovar is one of the tools included in the packages and generates a covariance matrix. The covariance matrix was used to generate a PCA plot in R.
Picard (*37*)	A set of Java command line tools for manipulating high-throughput sequencing data and formats. The two functions AddOrReplaceReadGroups and MarkDuplicates were used in the filtration process.
Plink (*45*)	Plink is a whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses. The focus of Plink is purely on analysis of genotype/phenotype data, and the program was used to generate a distance matrix used in a nexus file.
R (*38*)	R is a free software environment for statistical computing and graphics.
SAMtools (*35, 36*)	SAMtools provides various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.

Table 2 - continued

Program name	Short description
SeqMonk (*39*)	SeqMonk is a program that enables the visualization and analysis of mapped sequence data. It was written for use with mapped next generation sequence data but can in theory be used for any dataset which can be expressed as a series of genomic positions. The program was used to visualize differences between individuals with various amounts of data.
SplitsTree4 (*46*)	SplitsTree4 is an application for computing unrooted phylogenetic networks from molecular sequenced data. A distance matrix was used as the input file and the program computed a neighbor-network.
Stacks' Process_radtags (*31*)	Stacks is a software pipeline for building loci from short-read sequences, such as those generated on the Illumina platform. Stacks was developed to work with restriction enzyme-based data, such as RADSeq, for the purpose of building genetic maps and conducting population genomics and phylogeography. Process_radtags examines the raw read, discards non-usable segments, separates each individual from the combined library, and stacks together the reads from the same individual.
Treemix (*48*)	TreeMix is a method for inferring the patterns of population splits and mixtures in the history of a set of populations. In the underlying model, the modern-day populations in a species are related to a common ancestor via a graph of ancestral populations. The program uses the allele frequencies in the modern populations to infer the structure of this graph.

Results

The quality and amount of data before and after filtering

The two raw RADSeq libraries (RAD2 and RAD3) represented 258.652.242 and 435.371.596 reads. After the process_radtags program in Stacks was done filtering and cleaning, the sample sizes were reduced to 121.334.440 reads (divided over 22 samples), and 308.640.649 reads (divided over 36 samples). The RAD2 library had 77.798.350 ambiguous barcodes, 23.937.976 low quality reads, and 35.581.476 ambiguous RAD-Tags that were removed (figure 4a and 4c). The RAD3 library had 85.573.966 ambiguous barcodes, 24.187.383 low quality reads, and 16.969.598 ambiguous RAD-Tags that were removed (figure 4b and 4d). The average number of reads in each sample was 5.515.202 and 8.573.351, respectively.

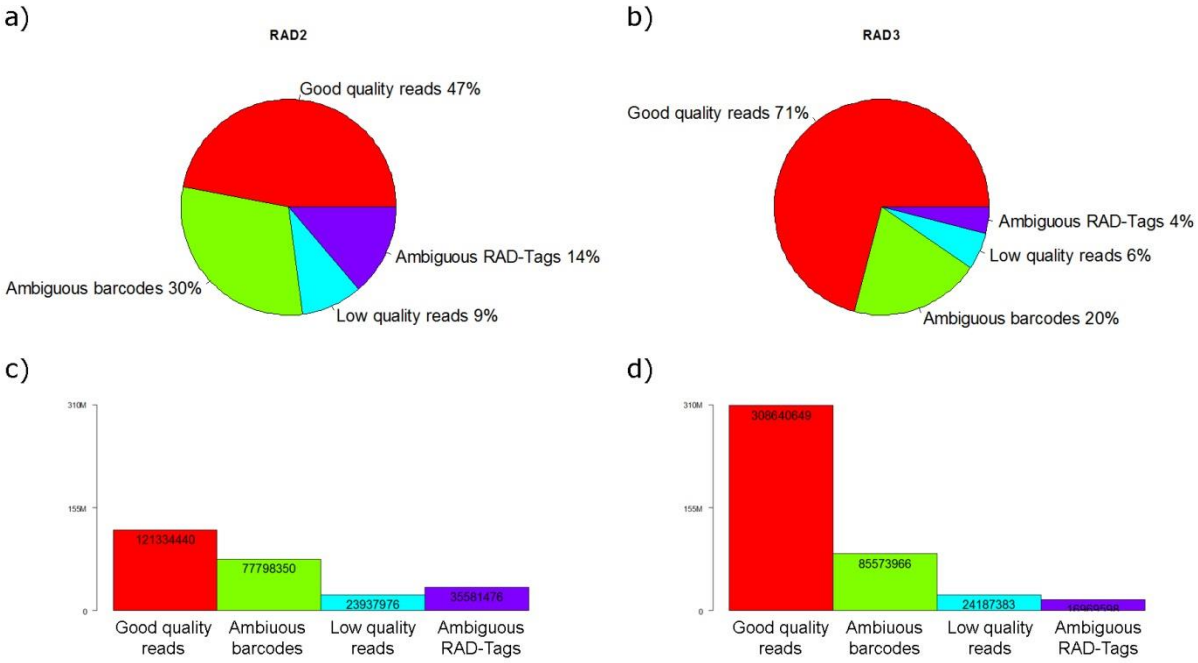


Figure 4 – Illustrations of quality and quantity of the two libraries after Stacks' process_radtags. Both the bar plots and pie charts illustrates the proportional differences within each library, but the bar plot also illustrates the quantitative differences between the two libraries.
a) The pie chart illustrates that the RAD2 library had 47% reads of good quality that were kept, and 30% ambiguous barcodes, 14% ambiguous RAD-Tags, and 9% reads of low quality that were removed.
b) The pie chart illustrates that the RAD3 library had 71% reads of good quality that were kept, and 20% ambiguous barcodes, 4% ambiguous RAD-Tags, and 6% reads of low quality that were removed.
c) The bar plot illustrates the differences in RAD2 between good quality reads (121.334.440), ambiguous barcodes (77.798.350), low quality reads (23.937.976) and ambiguous RAD-Tags (35.581.476).
d) The bar plot illustrates the differences in RAD3 between good quality reads (308.640.649), ambiguous barcodes (85.573.966), low quality reads (24.187.383) and ambiguous RAD-Tags (16.969.598).

The DNA quantification test revealed that the DNA concentration of the samples varied between 7 and 1000 µg/ml (table 1). Since the distribution of number of reads in the samples is skewed, I log-transformed the data. There was a small difference between the DNA quantification in the two libraries with a slightly higher DNA concentration in the second library (p-value 0.0493, figure 5a), and a linear model shows the correlation between the amount of DNA and number of reads per sample (figure 5b). The model for the linear regression gave a non-significant p-value of 0.5182 which is more than ten times higher the standard significance value of 0.05.

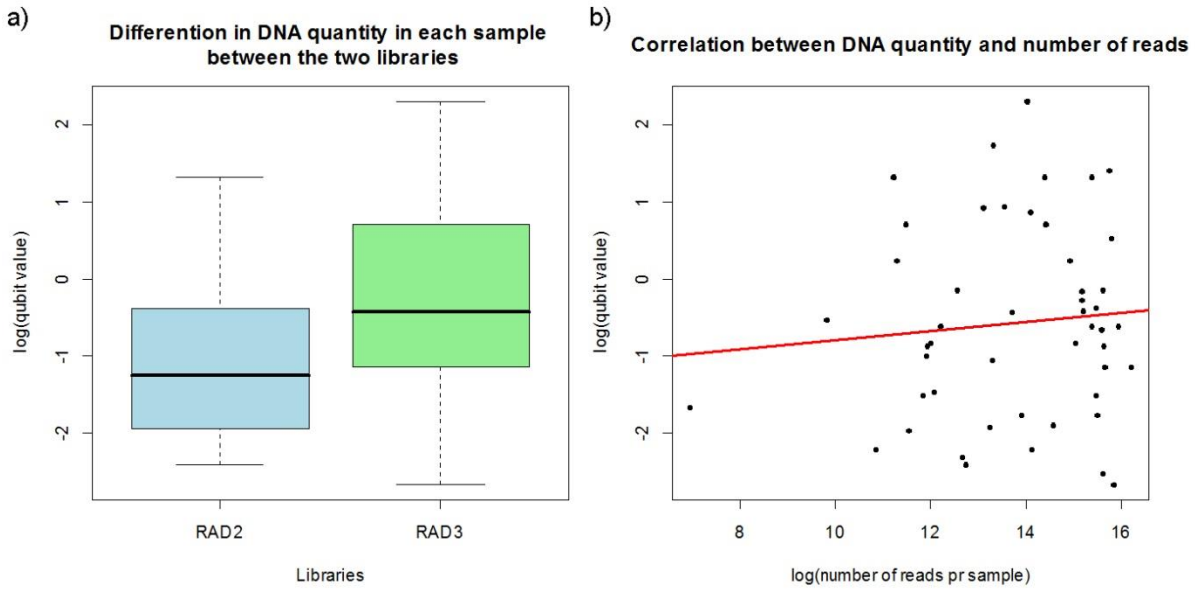


Figure 5 – Illustrations of DNA concentration in both libraries and in each sample.
a) The box plot illustrates the different log transformed DNA quantity in the two libraries. RAD3 has higher mean amount of DNA compared to RAD2 (p-value = 0.0493). RAD3 shows a higher variability.
b) The scatter plot shows the log-transformed correlation between the concentrations of DNA in each sample modeled by the number of reads after filtration. The red line is a linear model of the data illustrating any possible connection between DNA concentration and number of reads. The p-value of the linear model is 0.5182.

The forward and reversed fastq files for each sample were used as input files in Bowtie 2 when aligning each sample to the reference dog genome. The average alignment rate was 73.13 ± 21.85 (mean ± SD) and the median was 81.3%. The re-sequenced "afw03-2" had an alignment rate of only 0.90% and the second lowest was "afw18" with 28.51% alignment. The two individuals with the highest alignment rate were "afw08" and "afw25" with 92.12% and 92.26% respectively (see supplementary table 1 for a complete list). Only the aligned reads were included in the downstream analyses.

Through eight steps, the "RAD_BAM.sh" script filtered through all SAM-files and removed reads based on two criteria: minimum mapping quality with the resulting orphan reads and PCR duplication. A median of the number of accepted and discarded reads showed that the degree of accepted reads in RAD2 was 7% compared to RAD3, which had 70% accepted reads. Low mapping quality was responsible for removing 51% and 20% reads respectively, and PCR duplication represented 42% and 9% respectively of the total DNA (figure 6). The average number of reads in each file was 240.197 (RAD2) and 3.896.991 (RAD3).

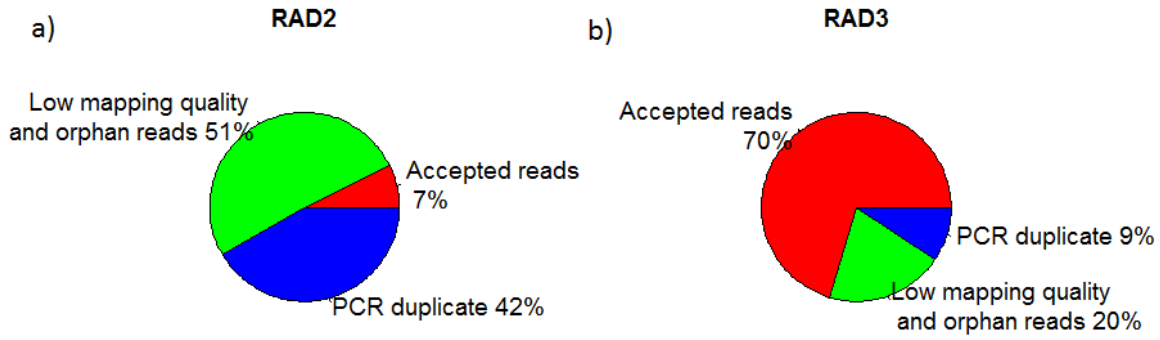


Figure 6 – Pie chart illustrating the median degree of accepted and discarded reads in the two libraries.
a) RAD2 consisted of 7% accepted reads, 42% PCR duplicates that were removed, and 51% low mapping quality with resulting orphan reads.
b) RAD3 consisted of 70% accepted reads, 9% PCR duplicates, and 20% low mapping quality with resulting orphan reads.

The amount of data was initially higher in RAD3 than in RAD2, but the most striking difference was in the quality of the collected data. For unknown reasons a higher proportion of reads were kept in all parts of the filtration process in the library with a new set of adapters. The reason why these adapters work differently is unknown.

Samples represented in both libraries were merged, and I excluded the files with too few reads. The enormous difference between the two libraries resulted in that the only samples kept from the first library were the ones that were also represented in the second library. After merging the samples represented in both libraries, no individual was only represented in RAD2. I wanted to just keep the best samples, but also of all relevant species. I had many more samples of African wolves available compared to the other species, and was therefore able to set a higher minimum number of how many reads the African wolf files had to contain in order to be accepted. Preferably I would like to have an equal number of individuals of all species, but without reducing the total number of individuals, this was not possible. Since African wolf was my main focus, and I looked for traces of hybridization in one of these individuals, I decided to keep 10 African wolves. I kept seven dogs and seven grey wolves since this was the maximum number I could use while having an equal number of samples of the two species. I only had four Ethiopian wolves available, and luckily they were all of good enough quality to be used. This left me with 28 individuals. The total number of reads in these files differed from 492.164 to 11.019.094 (figure 7).

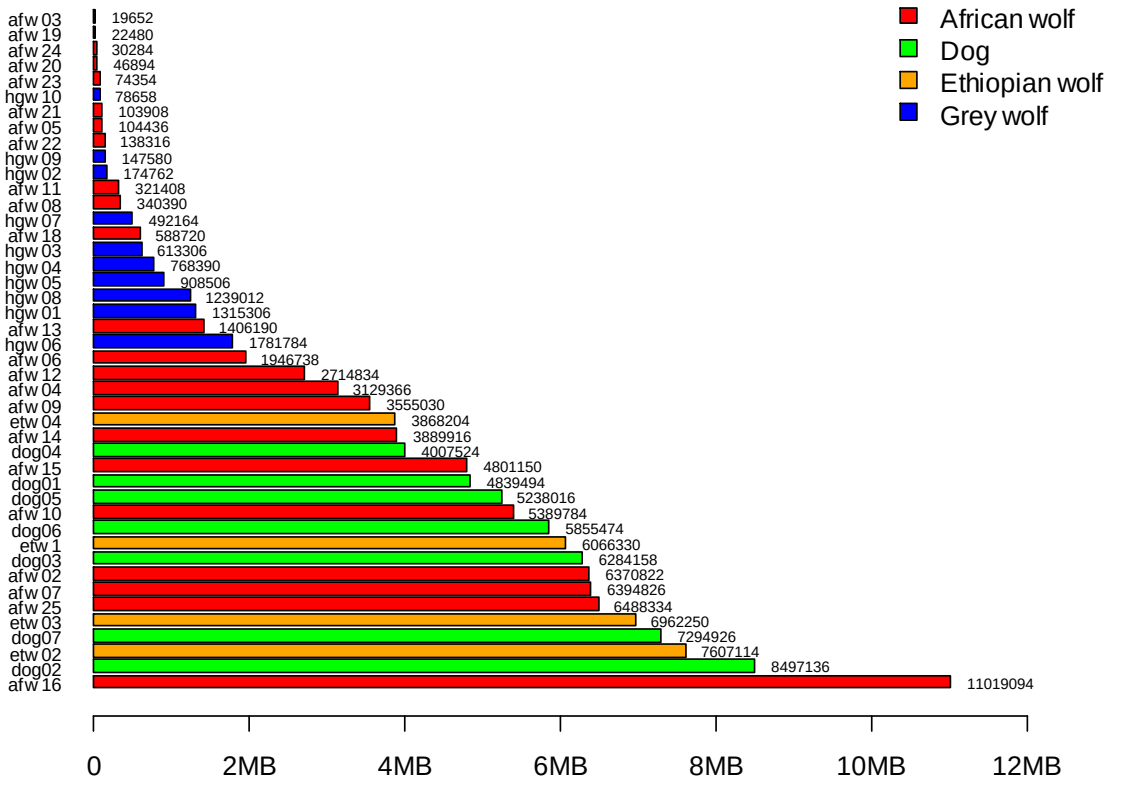


Figure 7 - Illustration of number of reads per sample after the filtration. All individuals of the relevant species are included, but some of them were omitted from the downstream analyses due to low sample size. “afw03” is the sample with the least number of reads (19.652 reads). “afw16” is the sample with the greatest number of reads (11.019.094 reads). The different species are marked with individual colors. African wolf = red, dog = green, Ethiopian wolf = orange, and grey wolf = blue.

Observing the filtered and aligned BAM files in SeqMonk illustrated how the reads are distributed in the genome (figure 8). Blue dots illustrate reads collected from the forward strand and red dots are collected from the reversed strand. The reads are spread out vertically where they would otherwise overlap. I included two files with quite different numbers of reads. “afw06.bam” contained 1.946.738 reads and “afw16.bam” contained 11.019.094. I selected a random position on a random chromosome and took a screenshot. The screenshot was manipulated in Photoshop CS6 in order to highlight the chromosome viewer and the content in each sample file.

(See supplementary figure 1 for the original screenshot.)

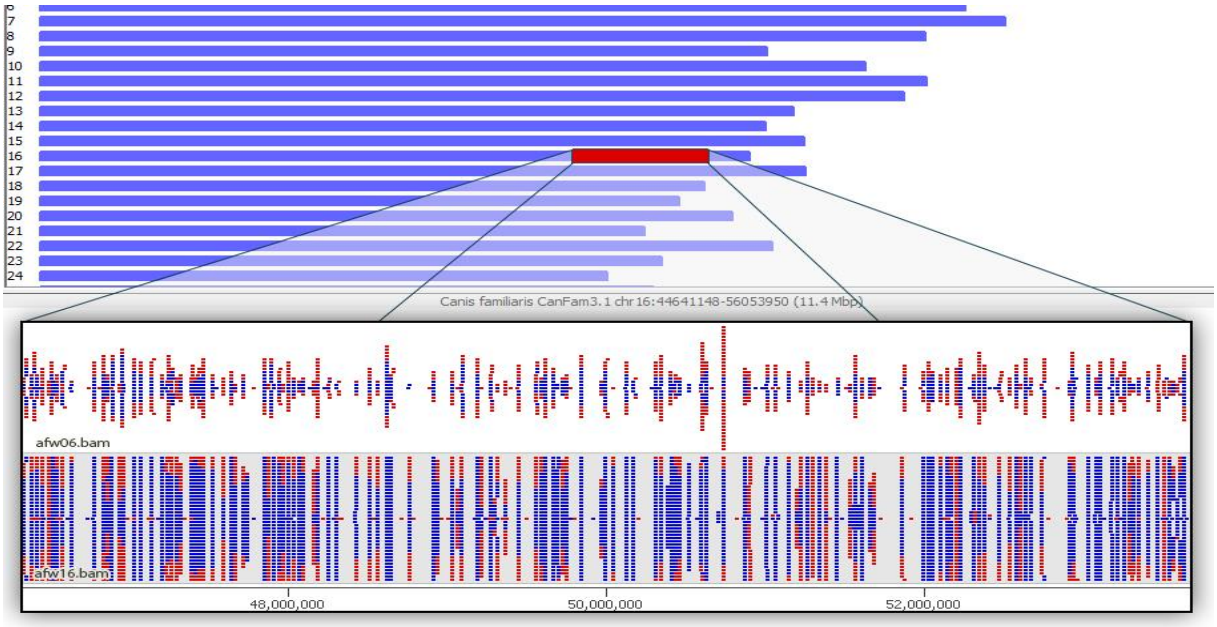


Figure 8 – Excepts from the screenshot image from SeqMonk illustrating the genome viewer and chromosome viewer. The reads are spread differently in the same position of the two individuals “afw06” (total number of reads = 1.946.738) and “afw16” (total number of reads = 11.019.094). All the reads are aligned to the dog genome and the selected part is chromosome 16, BP 44.641.148 - 56.053.950 (~11.4 million BP). Blue dots indicate forward reads, red dots indicate reversed reads.

The results of the population genomic analyses

All selected BAM files were included in ANGSD when I collected polymorphic sites in order to define similarities and differences between the individuals. I only collected SNPs with a frequency p-value less than 0.000001 (1e-6). ANGSD was run three times, producing new output files for different uses. In all three runs I received one more or less identical file. This was the major and minor allele frequencies calculation (.maf, table 2). The .maf files were not directly used in the downstream analysis, but were necessary to produce in order to do the correct calculations in the files that were used in the downstream analysis. Each run in ANGSD took around 16 CPU hours to run. Because it is possible to run the program on several threads at once, the wall time for each run was approximately three hours.

The other files produced by ANGSD used in downstream analysis, were a genotype calling file (.geno, table 3), an alternative genotype calling (.beagle, table 4), and a file that shows the frequencies of each different base (.counts, table 5).

Table 2 – An excerpt from the major and minor allele frequencies output from ANGSD (.maf) containing chromosome number (chromo), position in the chromosome (position), major (major) and minor (minor) allele for the specified individual (nInd), the reference's major allele (anc), and the likelihood in case of a known major and unknown minor (unknownEM).

chromo	position	major	minor	anc	unknownEM	nInd
10	90071	C	A	C	0.091066	28
10	92705	C	T	C	0.019843	27
10	104143	T	G	T	0.830939	8

Table 3 – Two excerpt from the genotype calling from ANGSD (.geno) listing chromosome (chromo) and position (position) for the two alternative alleles (allele1 and allele2). The continuing columns are the called genotype for each individual (ind01, ind02, …).
The first three example lines are from the genotype file used in PCA. The next three example lines are from the genotype file used in VCF where the genotypes are displayed as -1,0,1,2.

chromo	position	allele1	allele2	ind01	ind02	ind03	ind04	…
1	14000202	G	A	GG	NN	NN	GA	...
1	14000873	G	A	GG	GG	GG	NN	...
1	14001018	T	C	NN	NN	NN	CC	...
10	81824	A	G	2	2	2	1	...
10	82252	T	A	2	2	2	2	...
10	82253	A	G	1	1	0	0	...

Table 4 – An excerpt from the alternative genotype calling from ANGSD (.beagle) listing marker (marker) and position (position) for alternative alleles (allele1 and allele2). The continuing three columns contain values that equal one per site for each individual. This is just a normalization of the genotype likelihoods in order to avoid underflow problems in the beagle software it does not mean that they are genotype probabilities.

marker	position	allele1	allele2	ind0	ind0	ind0	Ind1	…
1	14000202	0	2	0.000000	0.001949	0.998051	0.000000	...
1	14000873	3	1	0.333333	0.333333	0.333333	0.333333	...
1	14001018	1	3	0.649432	0.324713	0.025854	0.666580	...

Table 5 – An excerpt from the count file listing the depth for each individual (.counts). This depth is the sum of reads covering a site for all individuals.

ind0TotDepth	ind1TotDepth	ind2TotDepth	ind3TotDepth	ind4TotDepth	….
9	4	1	6	3	...
0	0	0	0	0	...
0	1	0	0	0	...

The first run in ANGSD gave me a genotype calling (.geno) based on 2.953.943 SNPs that was used to generate a PCA (example table 3, alternative 1). The results from the first PCA (figure 9) revealed a clear difference between three of the supposed African wolves ("afw12", "afw14", and "afw15") and the other individuals. African wolf and grey wolf form a combined cluster, with high internal variability within the African wolf cluster, and a low internal variability within the grey wolf cluster. One of the African wolves, "afw25", is observed outside the 95% confidence interval ellipse, and is found in the midpoint between the semi-domestic dogs, and the rest of the African wolves. The Ethiopian wolves and the semi-domestic dogs form well-defined clusters with no overlap to the African wolf cluster and the grey wolf cluster. The clearest split is observed between the Ethiopian wolf cluster and the other species.

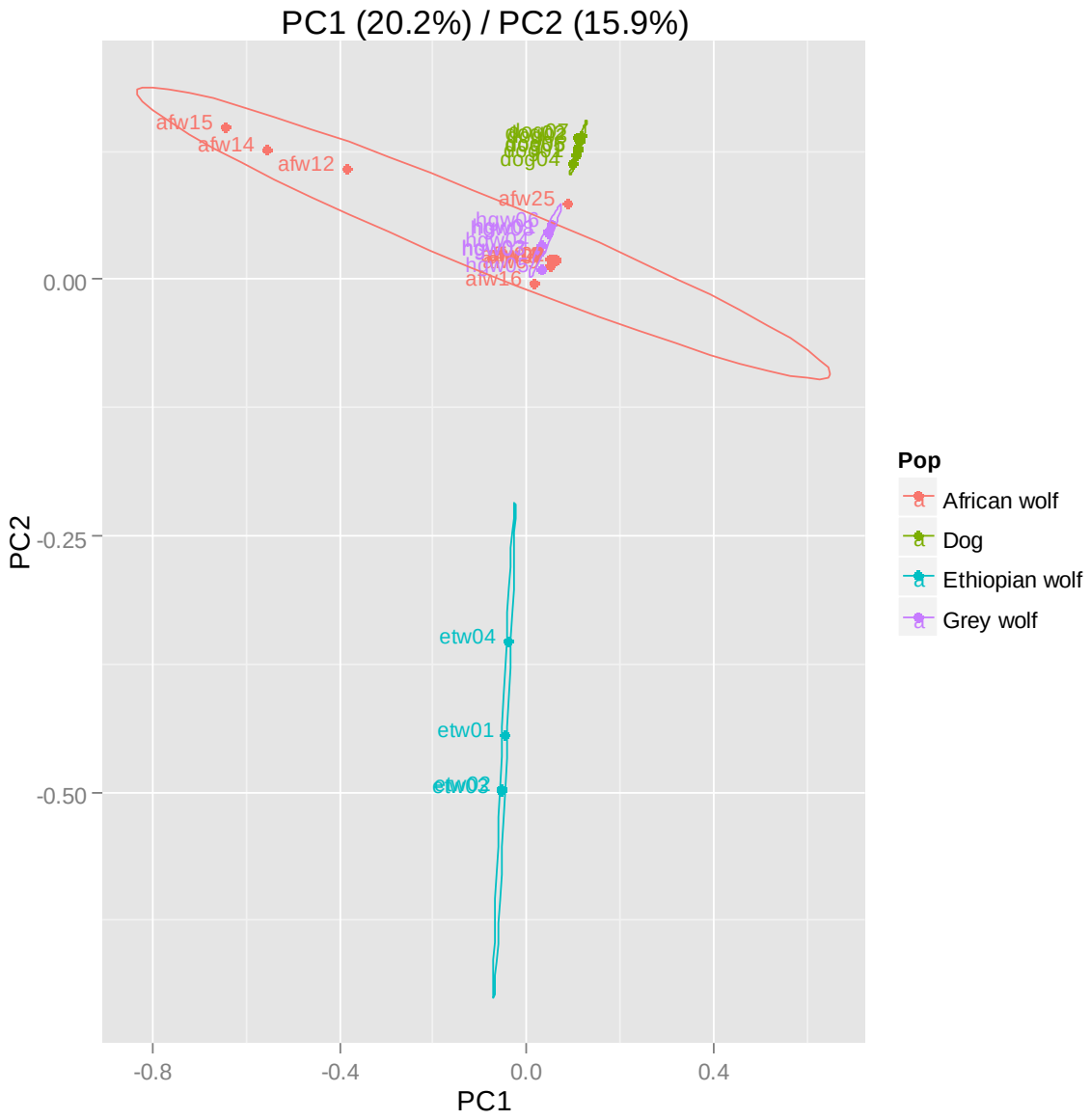


Figure 9 - Principal Component Analysis. The two first dimensions explain 20.2% and 15.9% respectively of the variation between the 28 included individuals. All species have a defined 95% confidence interval ellipse. The dog and the Ethiopian wolf forms separated clusters with no overlap with grey wolf or African wolf. The latter two forms a combined cluster with a high degree of internal variability within the African wolves, and a low internal variability within the grey wolves. The greatest split found in the first dimension, PC1, is between three of the African wolves ("afw12", "afw14", and "afw15") and all the other individuals. In addition to these three individuals, "afw25" is found outside the confidence interval ellipse, between the rest of the African wolves and dogs. The second dimension, PC2, explains a distinct split between the Ethiopian wolf and the other individuals. The internal variation in the Ethiopian wolf cluster is bigger than the internal variation in both dog and grey wolf clusters.

To confirm the possibility that three of the African wolves could be misidentified, I made an admixture plot with four populations (figure 10). This was based on a new run in ANGSD, generating a special genotype file (beagle format (table 4), 1.589.332 SNPs). The same three individuals ("afw12", "afw13", and "afw15") define a separate cluster. This plot also indicates that the individual marked "afw25" seems to be admixed equally between African wolf and semi-domestic dog.

Since "afw12", "afw14", and "afw15" made a separate cluster, and I only allowed four populations, grey wolves and semi-domestic dogs form a combined cluster. The grey wolf cluster shows a higher degree of admixture with some contribution from African wolves. The individual labeled "hgw05" shows some degree of contribution from Ethiopian wolf. The Ethiopian wolf and dog clusters show no degree of contribution from any other cluster.

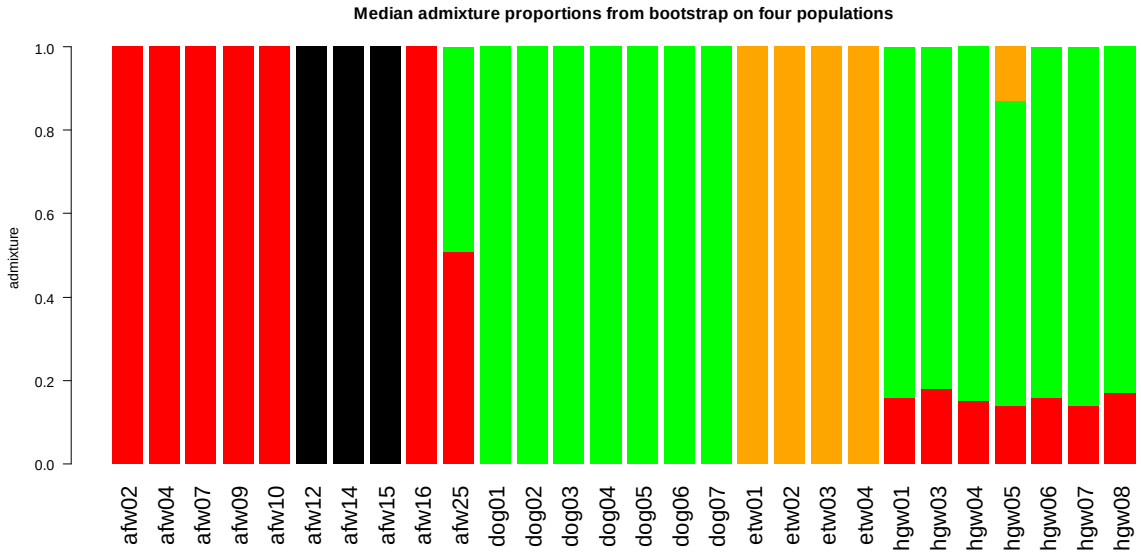


Figure 10 – This admixture plot illustrates the median admixture proportion of each individual based on 50 bootstrap replicates. The number of populations is set to four. The individuals "afw12", "afw14", and "afw15" define a separate cluster (black). The four Ethiopian wolves, "etw01", "etw02", "etw03", and "etw04", form a distinct cluster (orange) and all dogs and grey wolves are roughly one group (green). The grey wolves have some contribution from the rest of the African wolves (red, not "afw12", "afw14", or "afw15"). "hgw05" has some contribution from the Ethiopian wolf cluster, and "afw25" seems to be half dog, half African wolf. This plot is based on the matrix in the supplementary table 2.

Comparing with previously sequenced data (*3*), the three African wolf samples “afw12”, “afw14”, and “afw15” were identified as side-striped jackals. A true identification of these individuals has subsequently been done by E. Rueness (unpublished). The side-striped jackal is not relevant for this thesis, and therefore I replaced the three individuals with three African wolves that were previously removed due to small sample size. The labels of the new individuals were “afw06”, “afw13”, and “afw18”. The numbers of reads in these files were 1.946.738, 1.406.190, and 588.720 respectively. After running the last parts of the pipeline again, producing a new PCA plot (based on genotype calling file with 2.127.714 SNPs, figure 11), and a new admixture plot (based on beagle genotype file with 1.145.310 SNPs, figure 12), I could confirm that the new individuals were of the same species as the other African wolves. Even though there is high internal variation in the African wolf cluster, all individuals forms well-defined clusters with no overlap with the other clusters. Only “afw25” is observed outside the species 95% confidence interval ellipse.

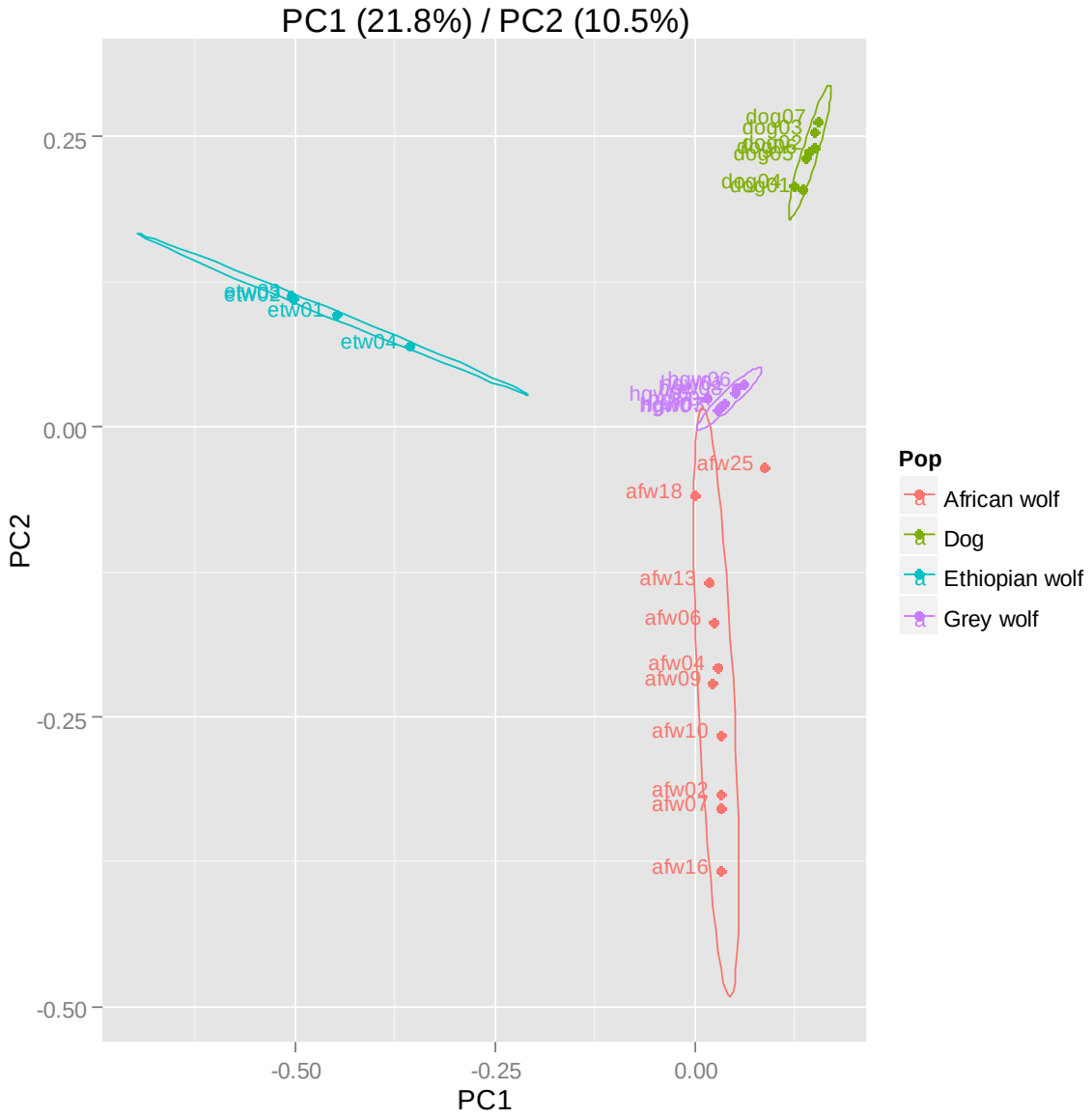


Figure 11 - Principal Component Analysis. The two first dimensions explain 21.8% and 10.5% respectively of the variation between the 28 included individuals. All species have a defined 95% confidence interval ellipse. All four species form distinct clusters with no overlap. The first dimension, PC1, explains a large difference between Ethiopian wolf and African wolf, grey wolf, and dog. In the second dimension, PC2, the largest split is observed between the dogs and the other species. The highest degree of internal variability is observed in the African wolf cluster, while the grey wolf is the cluster with least variation. The Ethiopian wolf shows a higher degree of internal variability than both grey wolf and dog. The African wolf “afw18” is placed slightly in the direction of the Ethiopian wolves, and “afw25” in the direction of the dogs and outside the confidence interval ellipse.

In the admixture analysis (figure 12) it still appears that “afw25” is almost equally mixed between African wolf and semi-domestic dog with some contribution from grey wolf (median admixture proportions are 0.46 (interquartile range (IQR) = 0.07) African wolf, 0.38 (IQR = 0.1) dog, and 0.16 (IQR = 0.17) grey wolf. See supplementary figure 3a). I can once more confirm that the three new individuals (“afw06”, “afw13”, and “afw18”) are correctly identified as African wolves. One of the new samples, “afw18”, has some contribution from Ethiopian wolf and grey wolf (median admixture proportions are 0.91 (IQR = 0.0575) African wolf, 0.07 (IQR = 0) Ethiopian wolf, and 0.02 (IQR = 0.06) grey wolf. See supplementary figure 3b).

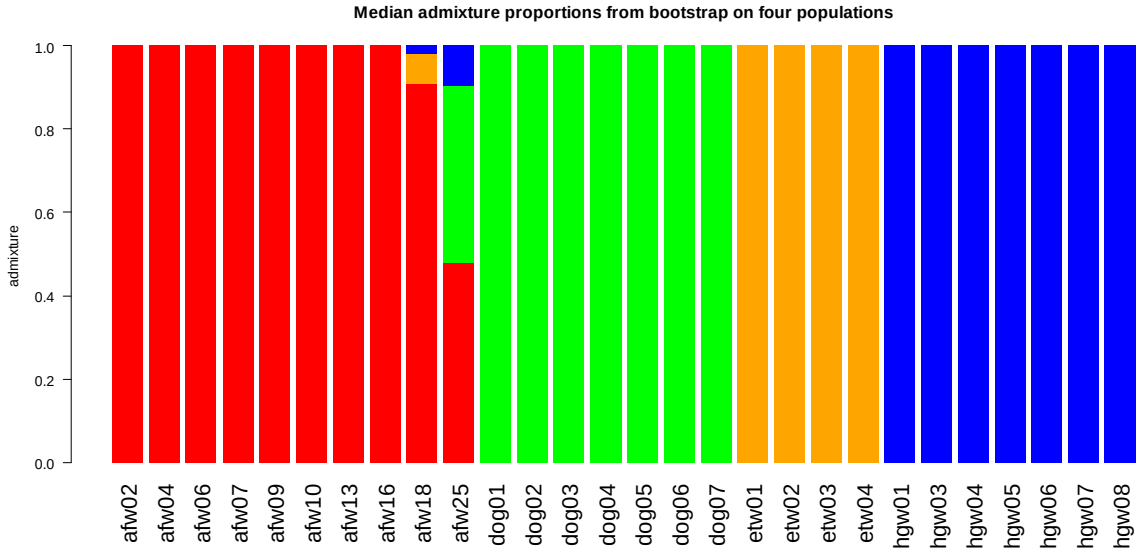


Figure 12 – Admixture plot illustrating the median admixture proportion in each individual based on 50 bootstrap replicates. The number of population is set to four. Each species form a distinct cluster, except “afw25” which seems to be an equal mix between African wolf and semi-domestic dog, with some contribution from grey wolf. “afw18” seems to have some contribution from Ethiopian wolf and grey wolf. This plot is based on the matrix in the supplementary table 3.

To confirm the results from the PCA (figure 9 and 11) and admixture analysis (figure 10 and 12) I checked for migration in TreeMix (figure 13). This was done through a converted VCF file (constructed by 1.145.309 SNPs from genotype calling (.geno) and base frequencies of each allele (.counts) from ANGSD). The TreeMix plot indicated the same results as those from the first admixture analysis (figure 10). The grey wolf cluster is influenced by gene flow from the Ethiopian wolf. The topology showed that dog and grey wolf are the most closely

related species, and combined they form a sister group to African wolf. Ethiopian wolf was specified to be the outgroup. The length of a branch along the x-axis indicates the uniqueness of the external node. The branch length of the Ethiopian wolf is undoubtedly the longest, and the dog branch is longer than both grey wolf and African wolf branches.

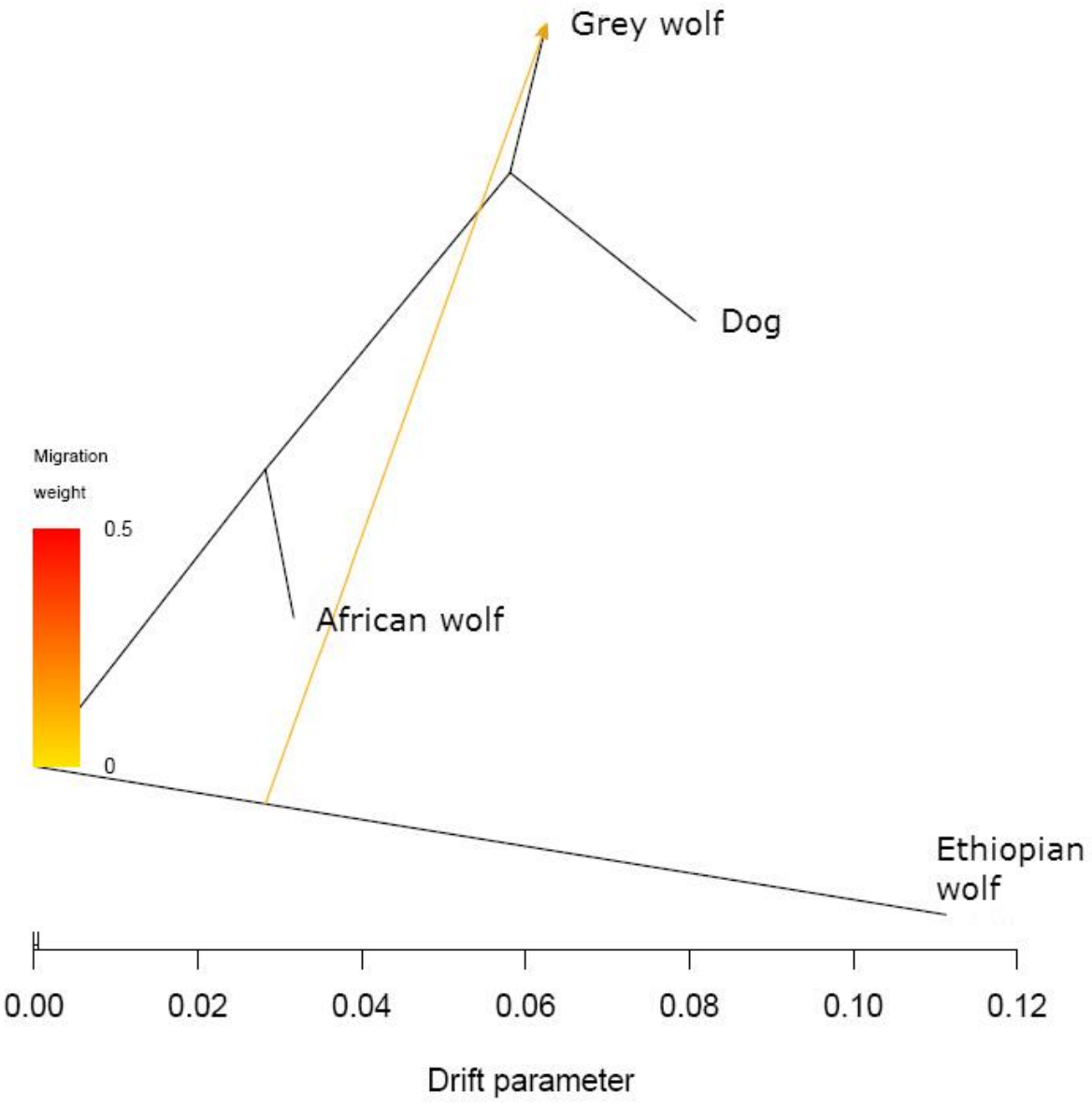


Figure 13 – Treemix4 generated a phylogenetic topology where I specified Ethiopian wolf as the outgroup. The monophyletic group of dogs and grey wolf defines a sister group to African wolf. The x-axis indicates the genetic drift. The length of the branches to each species indicates the degree of genetic drift from the last common ancestor. The arrow from Ethiopian wolf towards grey wolf indicates some degree of gene flow from the Ethiopian wolf in the grey wolf cluster. High migration weight is colored with red while low migration weight is colored with yellow. The arrow in this plot is orange, indicating a medium migration weight.

The results from the statistical program F4 showed “hgw05” was the individual that had the highest support for introgression. The mean ± SD of all the observed F4 values after downweighting was ~ 0.00534 ± 0.00052. The F4 value of “hgw05” was ~ 0.00324 (figure 14). The individual with the second highest contribution to the total introgression was “afw25” with a F4 value of ~ 0.00403 (see supplementary table 4 for complete list). The reason why “hgw05” is more extreme is probably because the program calculates each value on population level. With ten African wolves compared to seven grey wolves, the African wolf hybrid will be more diluted.

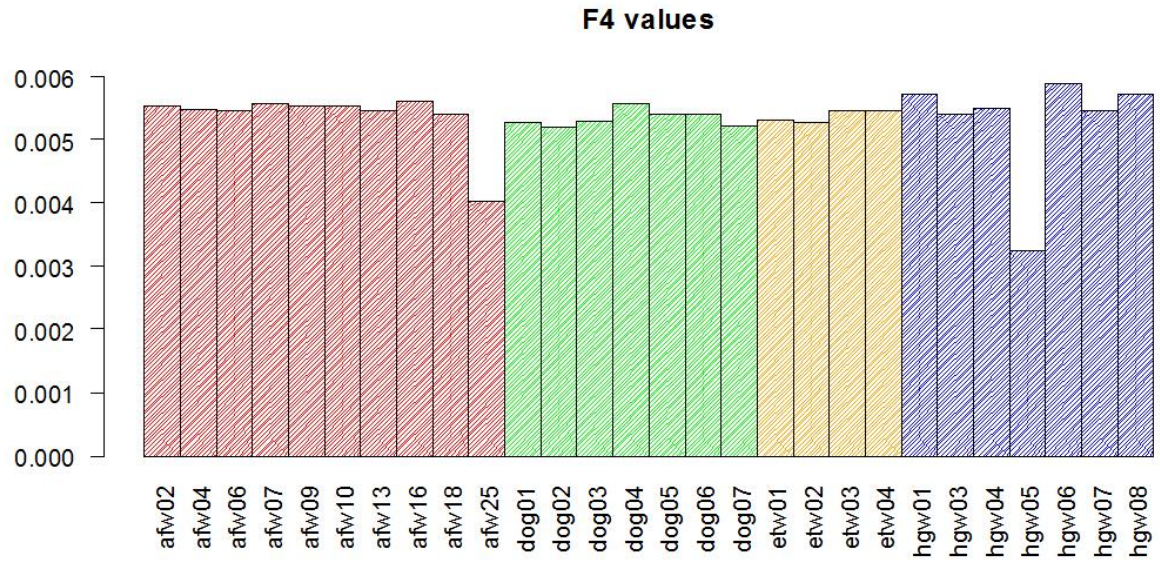


Figure 14 – Illustration of the F4 value for each sample after downweighting. The mean ± SD value is 0.005336071 ± 0.0005156107. “afw25” and “hgw05” stands out as the individuals with the highest impact on the overall introgression with F4 values of ~0.00403 and ~0.00324 respectively.

Based on the same VCF file used in TreeMix and F4, a new converted distance matrix was used in SplitsTree4. The neighbor network illustration from SplitsTree4 showed the relationships among the individuals and how they cluster according to species (figure 15). The most obvious split is between the four Ethiopian wolves and the rest of the samples. The length of the branch indicates a uniqueness of the cluster or individual that is not observed in the other individuals. There is a low degree of variation within the Ethiopian wolf cluster. The three other species form defined clusters, with the exception of three individuals, “afw18”, “afw25”, and “hgw05”. “afw18” is observed between the Ethiopian wolf and

African wolf clusters, “afw25” is observed between the dog and African wolf clusters, and “hgw05” is observed closer to the Ethiopian wolf cluster than the other grey wolves. Most of the African wolves are from Senegal (i.e. “afw06”, “afw16”, “afw02”, and “afw07”) and these individuals are slightly differentiated from the other African wolves. The dogs are also separated into two geographical clusters with “dog01” and “dog02” from Senegal and the last five from Ethiopia.

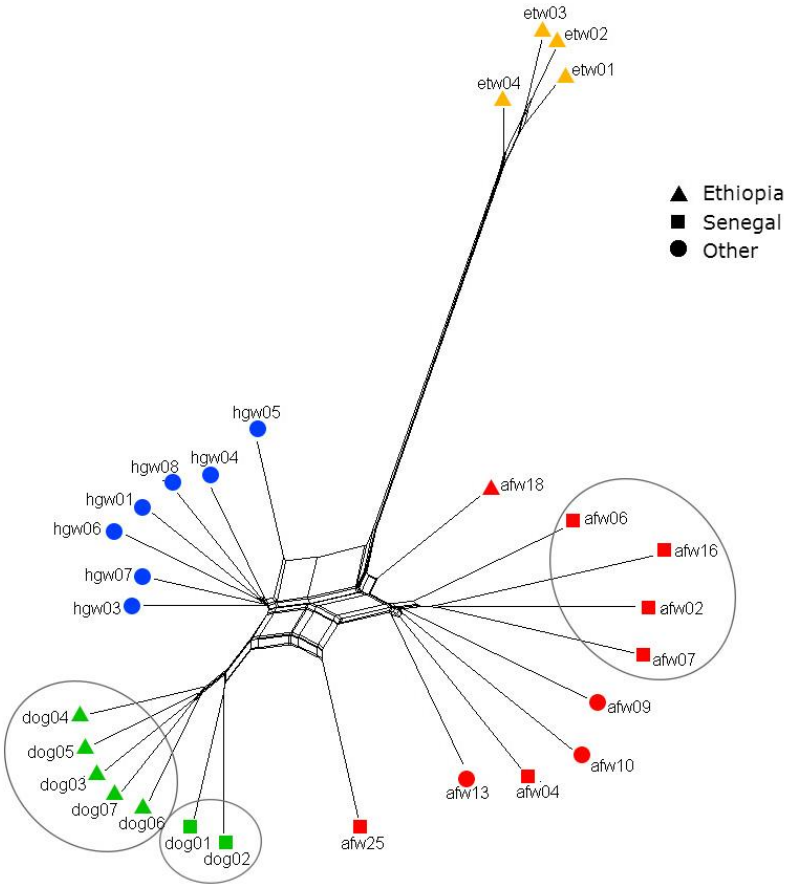


Figure 15 – Neighbor network generated in SplitsTree4 illustrating the relationship between the 28 samples. A clear split is observed between the four Ethiopian wolves and all the other samples. With the exception of three individuals, “afw18”, “afw25”, and “hgw05”, the three other species form defined clusters. “afw18” is observed between Ethiopian wolf and African wolf, “afw25” is observed between dog and African wolf, and “hgw05” is observed closer to the Ethiopian wolf cluster. Samples from Senegal are marked with a small square, samples from Ethiopia are marked with a small triangle, and all the other samples are marked with a circle. Four of the African wolves from Senegal (“afw06”, “afw16”, “afw02”, and “afw07”) are slightly differentiated from the other African wolves. “afw13” and “afw10” are from Algeria, “afw09” is from Mali, “afw18” is from Ethiopia, and “afw04” and “afw25” are from Senegal. A geographical split is also observed inside the dog cluster with “dog01” and “dog02” from Senegal, and the other five from Ethiopia.

Discussion

The aim of my study was to see if I could detect hybridization between African wolf and its sympatric canids. I used RADSeq data where I filtered out millions of SNPs in order to analyze them with a range of different bioinformatics tools. The resulting phylogenetic relationship among the species in my study is consistent with previous published data (*2, 3*), indicating that the choice of method was suitable to address my research question. Even the indication of geographical differentiations between populations in west and east Africa is consistent with some of the newest publications on the field (*2*).

RADSeq is proven to be a good way to obtain large amounts of data through several different studies (*3, 52-55*), including this thesis. Even though I only used RADSeq data to find traces of hybridization, other studies include everything from fine mapping of pond snail left-right asymmetry (*56*) to genetic diversity in beetle populations along a pollination gradient (*57*). However, RADSeq can be limiting when looking at individual genes because the fragments are short. This can be a problem when trying to separate incomplete lineage sorting from introgression. While other comparable studies have used D-statistic to test for introgression (*52, 58*), I used a program called F4 (49). This program has never been used in any published studies but has the advantage that it tests the individuals in the comparable populations, and not the population as a unit. However, because of the short fragments, and only selecting the most variable RAD loci, using RADSeq data poses a risk of introducing potential biases (*52*). Improvements in next-generation sequencing methods have potential to greatly improve the utility of RADSeq. Since the RADSeq protocols select fragment size performed by random shearing, even greater sequence lengths are achievable. Long contigs can be assembled from partly overlapping sheared-end reads resulting in up to several hundred BP in length (*59*). A “contig” (from contiguous) is a set of overlapping DNA segments that together represent a consensus region of DNA (*60*).

The program ANGSD (40), which I used to variant calling, was published in 2014, so it has not been on the marked for long. Still, it has already been used in a few recently published papers (*58, 61*). The advantages of implementation in downstream analysis have been exploited in both studies. In the study of Meyer et al. (*61*), looking at the evolutionary history of the blue-eyed black lemur, several of the downstream analyses I used in this thesis were also included. The study included an admixture plot and a PCA created with the same tools as I used (ngsAdmix (44) and ngsPopGen (43)), and they provided satisfying results. The study of Burri et al. (*58*), looking at linked selection and the recombination rate of *Ficedula* flycatchers, did not use the same downstream analysis, but used some of the same settings in ANGSD as in this thesis, such as genotype likelihoods and allele frequencies. They used some of the filtration features that ANGSD provides, which I excluded since I did the same filtration in other programs prior to ANSGD. Filtering the data prior to ANGSD or within ANGSD would not change the result, but since I had to run ANGSD several times, I saved a lot of time having already completed this part of the pipeline.

Using ANGSD in a study like this provides a lot of advantages, but also some drawbacks. Currently the results provided by ANGSD seem to be very good and suitable. Due to the short period it has been on the market, however, it may be too early to say whether the program is as stable as it seems. Another drawback with new software is that it can be more difficult to find solutions to technical problems, compared to a more widely used program. If I had challenges with SAMtools, for example, which is a very commonly used program, it was easy to find users that had experienced similar problems and already solved them. In ANGSD I had to depend on their wiki-page that functions as a manual. The manual is not very well organized, and it can be difficult to figure out which settings depend on what. Finding the right settings can be a time consuming and frustrating process. Still, I found the advantages of ANGSD to be more valuable than the drawbacks, and the result seems to be correct when compared to publications using the same methods or with similar topics.

The results from this study provide strong evidence of hybridization/gene flow between closely related canids in Africa. Although hybridization has been detected on several occasions in the *Canis* genus (*7, 8, 21-27*), it has never before been documented between African wolf and any sympatric canids. The two hybridization events found in this thesis are

located in Senegal and Ethiopia. In both locations the hybridizing species live sympatrically (*4*). In Senegal I found an individual that seemed to be a first generation hybrid between African wolf and dog, and in Ethiopia the hybridized individual was likely the result of an earlier cross-breeding event between Ethiopian wolf and African wolf, which backcrossed into African wolves. I had ten individuals from Ethiopia in this study (five dogs, four Ethiopian wolves and one African wolf) and none of the Ethiopian wolves or dogs seemed to be affected by any other species. The Ethiopian wolf is an endangered species and therefore closely monitored. According to the IUCN Red List (*5*), the biggest threat for this species is habitat loss and disease transmission from dogs.

If we assume that these results are correct, and African wolves hybridize with sympatric dogs and Ethiopian wolves, it can have a huge impact on the evolutionary development of all the involved species. Speciation reversal is the situation were two or more species morph together (*62*), and has been documented in at least two fish populations (*63, 64*). In both fish populations, the reversed speciation happened between sympatric subspecies. The relationship between these species could be closer compared to the relationships between African wolf and its sympatric canids. However, it has been speculated that extinction by hybridization and introgression is more important than commonly known in several taxa (*16*). Some endangered mammals are “contaminated” by hybridization and introgression, for example the Florida panther (*65*). But it does not mean that hybridization is the cause of the endangered status. In fact the hybrid panther kittens (i.e. those with a Texas ancestor) shows a higher survival rate than the pure bred Florida panther, and the adult pure bred female shows a higher mortality than the hybrid (*66*). This example is a case of documented inbreeding depression, and the advantageous outcome of hybridization may not be observed in other populations. In the case of hybridization between a male dog and respectively a female wolf or a female coyote, the lack of parental care from the dog will affect the survival of the cubs (*23*). Hybrid cubs may not be integrated in packs leaving them as only a reproductive waste. The burden of producing such progeny, may threaten a small population with extinction.

If an individual is a first generation hybrid, it means that there is not just a historical gene flow between these species, but an ongoing one. To be more certain about the number of

generations since the last hybridization event, it may be necessary to look at the recombination pattern. This could be done by visualizing the species-specific alleles from the two parental species in the hybrid individual, and see how large the recombinant parts are. If the individual is a first generation, we could expect quite large coherent parts from the same parental species. If the individual is a result of introgression, we would expect the recombinant parts to be smaller and more fragmented (*67*). I did not have the time to conduct that analysis, unfortunately.

It can be difficult to manage species if the boundaries between them become weaker. When individuals lose their species-specific character, it can in some cases be difficult for humans to identify the right species by phenotype. An example of misidentification is the three side-striped jackals first included in this study: "afw12", "afw14", and "afw15" which were collected as African wolves by experts in Ethiopia. However, for unknown reasons, the side-striped jackals in Ethiopia lack the characteristic side-stripe. Due to the large phenotypic variation in the African wolf and absence of characteristic phenotypic traits in the side-striped jackal, it was impossible to distinguish the two species. If it is necessary to DNA-test the individuals in order to distinguish them, the managing will be a much more demanding process.

Results from this study indicated hybridization between grey wolf and Ethiopian wolf, which is not possible since the species inhabit two different continents (North America and Africa). African wolf and Holarctic grey wolf are not sympatric species. But some mitochondrial studies indicate that Ethiopian wolf and coyote are closely related (*1, 2*). Even though my study is based on nuclear DNA, my results support that theory, a close relationship between Ethiopian wolf and coyote could be a logical explanation. Since hybridization between grey wolf and coyote is a common phenomenon (*7*), it is not surprising to find traces of coyote in one of the grey wolf samples.

At this point, with these results, it is impossible to determine what is causing the hybridization and how it will affect the populations. It is necessary to know more about how common these events are and if they appear more often in some areas. We do not know if the degree of gene flow is stable, increasing or decreasing between the species. A large-scale study on the frequency of hybridization between African canids over many years could tell

us, and with that knowledge it could be possible to estimate some aspects of these species' future. A similar study was done with the tree-spined stickleback when detecting the speciation reversal (*63*), but it would be much more challenging to conduct a study like that on a large mammal that moves across large geographical areas. However, learning more about these canids and their relationships to each other is very helpful when developing conservation guidelines for each species.

It would also be worth looking into how hybridization affects the behavior of an individual. If changes in behavior result in conflicts with humans, that will have an important impact on the management of the species. Since this is the first documented event of hybridization between African wolf and its sympatric canids, no trends in behavior changes in African wolf hybrids have been reported.

Conclusion

Next-generation sequencing and RADSeq are ideal ways to find signs of introgression and hybridization within the genus *Canis*. The phylogenetic topology of the species came out as expected compared with earlier studies, also displaying geographical variation. For the first time has hybridization between African wolf and sympatric canids been detected and confirmed. I found two hybridization events; the first was a possible first generation between African wolf and semi-domestic dog. This individual was collected in Senegal, where both species live sympatric. The second event was an African wolf with traces of Ethiopian wolf. This individual was collected in Ethiopia, the only place where both species exist.

Of the 10 African wolves included in my study, two of them showed signs of hybridization, which is a large proportion. But 10 individuals are not enough to conclude how common these events are. More research is needed to gain a more comprehensive picture.

References

1. Rueness, EK *et al.* (2011) The Cryptic African Wolf: *Canis aureus lupaster* Is Not a Golden Jackal and Is Not Endemic to Egypt. *PloS ONE,* **6**, doi:10.1371/journal.pone.0016385.

2. Koepfli, KP *et al.* (2015) Genome-wide Evidence Reveals that African and Eurasian Golden Jackals Are Distinct Species. *Current Biology,* **25**, 2158–2165.

3. Rueness, EK *et al.* (2015) The African Wolf is a Missing Link in the Wolf-like Canid Phylogeny. *BioRxiv,* doi:http://dx.doi.org/10.1101/017996.

4. Gaubert, P *et al.* (2012) Reviving the African Wolf *Canis lupus lupaster* in North and West Africa: A Mitochondrial Lineage Ranging More than 6,000 km Wide. *PLoS ONE,* **7**, doi:10.1371/journal.pone.0042740.

5. Marino J, Sillero-Zubiri C. (2013) *Canis simensis*. *The IUCN Red List of Threatened Species 2013,* e.T3748A10051312.

6. vonHoldt BM *et al. (2011)* A Genome-wide Perspective on the Evolutionary History of Enigmatic Wolf-like Canids. *Genome Research,* **21**, 1294-1305.

7. Lehman *N et al. (1991)* Introgression of Coyote Mitochondrial DNA Into Sympatric North American Gray Wolf Populations. *Evolution,* **45**, 104-119.

8. Gottelli D *et al*. (1994) Molecular Genetics of the Most Endangered Canid: the Ethiopian wolf *Canis simensis*. *Molecular Ecology,* **3**, 301-312.

9. Randi E, Lucchini V (2002) Detecting rare Introgression of Domestic Dog Genes into Wild Wolf (*Canis lupus*) Populations by Bayesian Admixture Analyses of Microsatellite Variation. *Conservation Genetics,* **3**, 29-43.

10. Hailer F, Leonard JA (2008) Hybridization among Three Native North American Canis Species in a Region of Natural Sympatry. *PLoS ONE,* **3**, doi:10.1371/journal.pone.0003333.

11. Twyford AD, Ennos RA (2011) Next Generation Hybridization and Introgression. *Heredity,* **108**, 179-189.

12. Grada A, Weinbrecht K (2013) Next-Generation Sequencing: Methodology and Application. *Journal of Investigative Dermatology,* **133**, doi:10.1038/jid.2013.248.

13. Bentley DR *et al.* (2008) Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry. *Nature,* **456**, 53-59.

14. Davey JW, Blaxter ML (2010) RADSeq: Next-Generation Population Genetics. *Functional Genomics,* **9**, 416-423.

15. Hogeweg P (2011) The Roots of Bioinformatics in Theoretical Biology. *PLoS Computational Biology,* **7**, doi:10.1371/journal.pcbi.1002021.

16. Rhymer JM, Simberloff D (1996) Extinction by Hybridization and Introgression. *Annual Review of Ecology and Systematics,* **27**, 83-109.

17. Rieseberg LH, Wendel JF (1993), Introgression and its Consequences in Plants. *Hybrid zones and the evolutionary process 01/1993, Oxford University Press*.

18. Rhymer JM, Williams MJ, Braun MJ (1994) Mitochondrial Analysis of Gene Flow Between New Zealand Mallards (*Anas platyrhynchos*) and Grey Duck (*A. super-ciliosa*). *The Auk,* **111**, 970-978.

19. Choleva L *et al.* (2014) Distinguishing between Incomplete Lineage Sorting and Genomic Introgression: Complete Fixation of Allospecific Mitochondrial DNA in a Sexually Reproducing Fish (*Cobitis*; *Teleostei*), despite Clonal Reproduction of Hybrids. *PLoS ONE,* **9**, doi:10.1371/journal.pone.0080641.

20. Rogers J, Gibbs RA (2014) Comperative Primate Genomics: Emerging Pattern of Genome Content and Dynamics. *Nature Reviews,* **15**, 347–359.

21. Iacolina L *et al.* (2010) Y-chromosome Microsatellite Variation in Italian Wolves: A Contribution to the Study of Wolf-Dog Hybridization Patterns. *Mammalian Biology,* **75**, 341–347.

22. Hindrikso M *et al.* (2012) Bucking the Trend in Wolf-Dog Hybridization: First Evidence from Europe of Hybridization between Female dogs and Male Wolves. *PLoS ONE,* **7**, doi:10.1371/journal.pone.0046465.

23. Vilà C, Wayne RK (1999) Hybridization Between Wolves and Dogs. *Conservation Biology,* **13**, 195-198.

24. Kopaliani N *et al.* (2014) Gene Flow Between Wolf and Shepherd Dog Populations in Georgia (Caucasus). *Journal of Hereditary,* **105**, doi: 10.1093/jhered/esu014.

25. Gondinho R *et al.* (2011) Genetic Evidence for Multiple Events of Hybridization between Wolves and Domestic Dogs in the Iberian Peninsula. *Molecular Ecology,* **20**, 5154-5166.

26. Munõz-Fuentes V *et al.* (2010) The Genetic Legacy of Extirpation and Re-Colonization in Vancouver Island Wolves. *Conservation Biology,* **11**, 547-556.

27. Hennelly L, Habib B, Lyngdoh S (2015) Himalayan Wolf and Feral Dog Displaying Mating Behavior in Spiti Valley, India, and Potential Conservation Threads from Sympatric Feral Dogs. *Canid Biology and Conservation,* **18**, 33-36.

28. Kelly BT, Beyer A, Phillips MK (2008) *Canis rufus*. *The IUCN Red List of Threatened Species 2008,* e.T3747A10057394.

29. Barton NH (2001) The role of Hybridization in Evolution. *Molecular Ecology,* **10**, 551-568.

30. Baird NA *et al.* (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE*, **10**, doi:10.1371/journal.pone.0003376.

31. Catchen J *et al.* (2013) Stacks: An Analysis Tool Set for Population Genomics. *Molecular Ecology,* **22**, 3124-3140.

32. Flicek P, Birney E (2009) Sense from Sequence Reads: Methods for Alignment and Assembly. *Nature Methods,* **6**, 6-12.

33. Langmead B, Salzberg SL (2012) Fast Gapped-Read Alignment with Bowtie 2. *Nature Methods,* **9**, 357-359.

34. Cunningham F *et al.* (2015) Ensembl 2015. *Nucleic Acids Research 2015,* **43**, doi:10.1093/nar/gku1010.

35. Li H *et al.* (2009) The Sequence Alignment/Map Format and SAMtools. *Bioinfomatics application note,* **25**, 2078-2079.

36. Li H, A (2011) Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data. *Bioinformatics,* **27**, 2987-2993.

37. Broadinstitute, Picard. http://broadinstitute.github.io/picard/

38. R Core Team (2015) A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria.* http://www.R-project.org

39. Andrews S (2015) SeqMonk, *Babraham Bioinformatics*. http://www.bioinformatics.babraham.ac.uk/projects/seqmonk

40. Korneliussen TS, Albrechtsen A, Nielsen R (2011) ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics,* **15**, doi:10.1186/s12859-014-0356-4.

41. Kim SY *et al.* (2011) Estimation of Allele Frequency and Association Mapping using Next-Generation Sequencing Data. *BMC Bioinformatics,* **12**, doi:10.1186/1471-2105-12-231

42. Nielsen R *et al.* (2012) SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. *PLoS ONE,* **7**. doi:10.1371/journal.pone.0037558.

43. Fumagalli M (2015) ngsPopGen. https://github.com/mfumagalli/ngsPopGen

44. Skotte L, Korneliussen T, Albrechtsen A (2013) Estimating Individual Admixture Proportions from Next Generation Sequencing Data. *Genetics,* **113**, 693-702.

45. Purcell S *et al.* (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics,* **81**, 559-575.

46. Huson DH, Bryant D (2006) Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology Evolution,* **23**, 254-267.

47. Bryant D, Moulton V (2004) Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. *Molecular Biology Evolution,* **21**, 255-265.

48. Pickrell JK, Pritchard JK (2012) Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genet,* **8**, doi:10.1371/journal.pgen.1002967.

49. Matschiner M (2015) F4. *Unpublished*.

50. Reich D *et al.* (2009) Reconstructing Indian Population History. *Nature,* **461**, 489-494.

51. Andrews S (2010) FastQC. *Babraham Bioinformatics*. http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc

52. Eaton DAR, Ree RH (2013) Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (Pedicularis: Orobanchaceae). *Systematic Biology,* **62**, 689-706.

53. Leaché AD *et al.* (2015) Phylogenomics of Phrynosomatid Lizards: Conflicting Signals from Sequence Capture versus Restriction Site Associated DNA Sequencing. *Genome Biology and Evolution,* **7**, doi:10.1093/gbe/evv026.

54. Rutledge LY *et al.* (2015) RAD Sequencing and Genomic Simulations resolve Hybrid Origins within North American Canis. *Biology Letters,* **11**, doi:10.1098/rsbl.2015.0303.

55. Combosch DJ, Vollmer SV (2015) Trans-Pacific RAD-Seq Population Genomics confirms Introgressive Hybridization in Eastern Pacific Pocillopora Corals. *Molecular Phylogenetics and Evolution,* **88**, 154-162.

56. Liu MM *et al.* (2013) Fine Mapping of the Pond Snail Left-Right Asymmetry (Chirality) Locus Using RAD-Seq and Fibre-FISH. *PLoS ONE,* **8**, doi:10.1371/journal.pone.0071067.

57. Giska I *et al.* (2015), Genome-wide Genetic Diversity of Rove Beetle Populations along a Metal Pollution Gradient. *Ecotoxicology and Environmental Safety,* **119**, 98-105.

58. Burri R *et al.* (2015) Linked Selection and Recombination Rate Variation drive the Evolution of the Genomic Landscape of Differentiation across the Speciation continuum of *Ficedula* flycatchers. *Genome Research,* **25**, doi:10.1101/gr.196485.115.

59. Etter P *et al.* (2011) Local *de nova* Assembly of RAD Pair-End Contigs using Short Sequencing Reads. *PLoS ONE,* **6**, doi:10.1371/journal.pone.0018561.

60. Gregory SG (2005) Contig Assembly. *eLS*, doi:10.1038/npg.els.0005365

61. Meyer WK *et al.* (2015) Evolutionary History Inferred from the *de novo* Assembly of a Nonmodel Organism, the Blue-Eyed Black Lemur. *Molecular Ecology,* **24**, 4392-4405.

62. Coyne JA, Orr HA (2004) *Speciation*. Sinauer Associates, Sunderland, Massachusetts. ISBN: 978-0-87893-089-0

63. Taylor EB *et al.* (2006) Speciation in Reverse: Morphological and Genetic Evidence of the Collapse of a Three-Spined Stickleback (*Gasterosteus aculeatus*) Species pair. *Molecular Ecology,* **15**, 343-355.

64. Bhat S *et al.* (2014) Speciation Reversal in European Whitefish (*Coregonus lavaretus* (L.)) Caused by Competitor Invasion. *PLoS ONE,* **9**, doi:10.1371/journal.pone.0091208.

65. Fergus C. (1991) The Florida Panther verges on Extinction. *Science,* **251**, 1178-1180.

66. Pimm SL, Dollar L, Bass Jr OL (2006) The Genetic Rescue of the Florida Panther. *Animal Conservation,* **9**, doi:10.1111/j.1469-1795.2005.00010.x.

67. Fu Q *et al.* (2015) An Early Modern Human from Romania with a Recent Neanderthal ancestor. *Nature,* **524**, 216–219.

Appendix

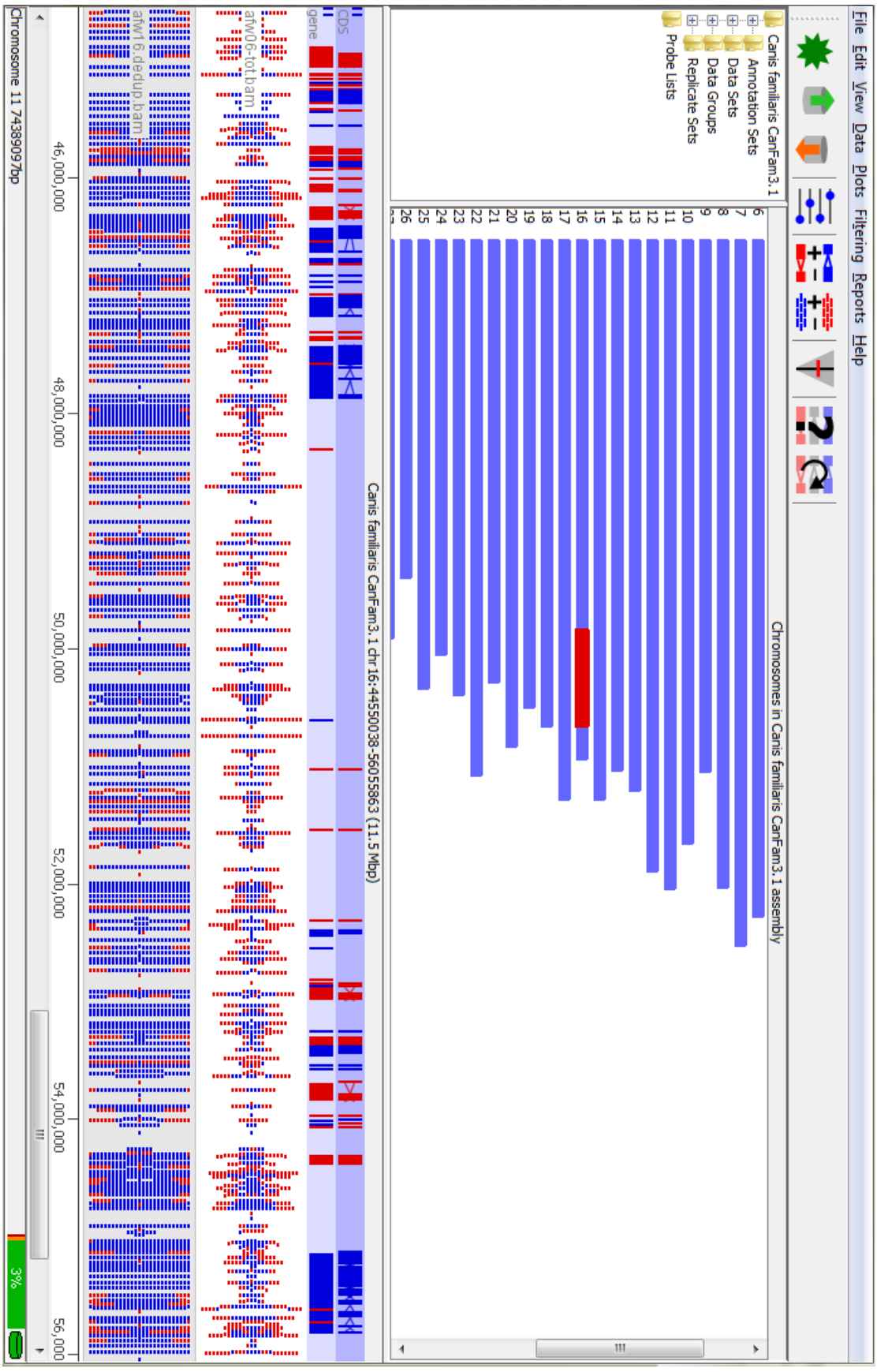
Supplementary table 1 – A list of Bowtie 2 results with number of reads in each input file and how many reads that aligned concordantly 0 times, exactly 1 time, and > 1 times. The last column lists overall alignment rate per sample.

Label	Number of reads	0	1	>1	Percent alignment
afw01	1432695	297657	1019695	115343	79.22%
afw02	1125568	104286	953400	67882	90.73%
afw02-2	5658332	842327	4474046	341959	85.11%
afw03	93045	85010	7117	918	8.64%
afw03-2	1962601	1944841	15950	1810	0.90%
afw04	725276	131803	545867	47606	81.83%
afw04-2	3627441	1351260	2086544	189637	62.75%
afw05	758448	73481	636514	48453	90.31%
afw06	712757	83160	584458	45139	88.33%
afw06-6	1842524	443924	1286211	112389	75.91%
afw07	2025948	198002	1705077	122869	90.23%
afw07-2	5731683	1142434	4279892	309357	80.07%
afw08	2398291	189070	2073543	135678	92.12%
afw09	1157828	105556	985050	67222	90.88%
afw09-2	3075037	551236	2357721	166080	82.07%
afw10	957507	83794	818739	54974	91.25%
afw10-1	4586649	618973	3725412	242264	86.50%
afw11	2258271	195949	1918611	143711	91.32%
afw12	4066100	342378	3516318	207404	91.58%
afw12-2	1882153	303102	1490103	88948	83.90%
afw13	429488	63867	327148	38473	85.13%
afw13-2	1787233	598343	1085086	103804	66.52%

Supplementary table 1 - continued

Label	Number of reads	0	1	>1	Percent alignment
afw14	3583119	846272	2557531	179316	76.38%
afw15	4121230	804176	3136116	180938	80.49%
afw16	13268084	2657783	9910727	699574	79.97%
afw18	2096407	1498762	523873	73772	28.51%
afw25	7483660	579534	6479581	424545	92.26%
afw25-2	4815734	685755	3876497	253482	85.76%
dog01	593561	81919	469563	42079	86.20%
dog01-1	4944637	1221789	3389092	333756	75.29%
dog02	1416660	153523	1185743	77394	89.16%
dog02-2	7435734	1228136	5811683	395915	83.48%
dog03	5171041	933265	3975168	262608	81.95%
dog04	3202655	570200	2462146	170309	82.20%
dog05	4354627	823000	3306733	224894	81.10%
dog06	5017611	965356	3774848	277407	80.76%
dog07	6205537	1160138	4718744	326655	81.30%
etw01	6940632	1110599	5472104	357929	84.00%
etw02	8706213	1471816	6750345	484052	83.09%
etw03	7102410	1970753	4636197	495460	72.25%
etw04	3410461	684226	2520619	205616	79.94%
hgw01	1390787	494737	810608	85442	64.43%
hgw02	320325	202254	101692	16379	36.86%
hgw03	1499249	1038675	402096	58478	30.72%
hgw04	1539449	824902	634933	79614	46.42%
hgw05	1618358	816233	705075	97050	49.56%
hgw06	3060078	1349332	1512823	197923	55.91%
hgw07	1154784	677590	423623	53571	41.32%
hgw08	1997081	825914	1056001	115166	58.64%

Figure 1 – Screenshot from SeqMonk before manipulating the image in order to highlight the content of each sample file



Supplementary table 2 – Admixture proportions of four clusters manually calculated median value from 50 bootstrap replicates on the first admixture run from ANGSD including side-striped jackals.

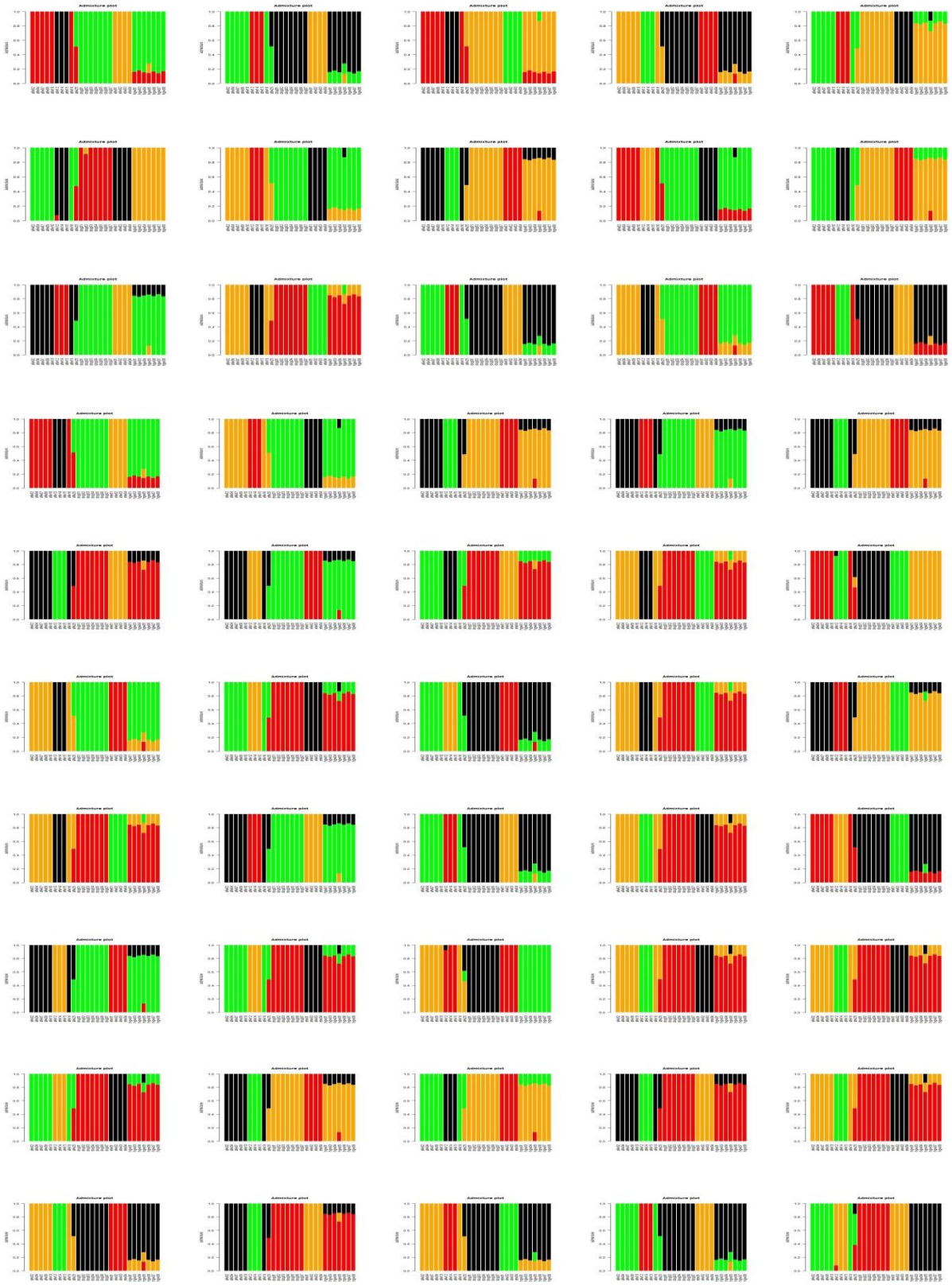
1.0000	0.0000	0.0000	0.0000
1.0000	0.0000	0.0000	0.0000
1.0000	0.0000	0.0000	0.0000
1.0000	0.0000	0.0000	0.0000
1.0000	0.0000	0.0000	0.0000
0.0000	1.0000	0.0000	0.0000
0.0000	1.0000	0.0000	0.0000
0.0000	1.0000	0.0000	0.0000
1.0000	0.0000	0.0000	0.0000
0.5100	0.0000	0.4900	0.0000
0.0000	0.0000	1.0000	0.0000
0.0000	0.0000	1.0000	0.0000
0.0000	0.0000	1.0000	0.0000
0.0000	0.0000	1.0000	0.0000
0.0000	0.0000	1.0000	0.0000
0.0000	0.0000	1.0000	0.0000
0.0000	0.0000	1.0000	0.0000
0.0000	0.0000	0.0000	1.0000
0.0000	0.0000	0.0000	1.0000
0.0000	0.0000	0.0000	1.0000
0.0000	0.0000	0.0000	1.0000
0.1600	0.0000	0.8400	0.0000
0.1800	0.0000	0.8200	0.0000
0.1500	0.0000	0.8500	0.0000
0.1400	0.0000	0.7300	0.1300
0.1600	0.0000	0.8400	0.0000
0.1400	0.0000	0.8600	0.0000
0.1700	0.0000	0.8300	0.0000

Supplementary table 3 – Admixture proportions of four clusters manually calculated median value from 50 bootstrap replicates from the second admixture run from ANGSD without side-striped jackals.

1.0000	0.0000	0.0000	0.0000
1.0000	0.0000	0.0000	0.0000
1.0000	0.0000	0.0000	0.0000
1.0000	0.0000	0.0000	0.0000
1.0000	0.0000	0.0000	0.0000
1.0000	0.0000	0.0000	0.0000
1.0000	0.0000	0.0000	0.0000
1.0000	0.0000	0.0000	0.0000
0.9100	0.0000	0.0700	0.0200
0.4806	0.4228	0.0000	0.0966
0.0000	1.0000	0.0000	0.0000
0.0000	1.0000	0.0000	0.0000
0.0000	1.0000	0.0000	0.0000
0.0000	1.0000	0.0000	0.0000
0.0000	1.0000	0.0000	0.0000
0.0000	1.0000	0.0000	0.0000
0.0000	1.0000	0.0000	0.0000
0.0000	0.0000	1.0000	0.0000
0.0000	0.0000	1.0000	0.0000
0.0000	0.0000	1.0000	0.0000
0.0000	0.0000	1.0000	0.0000
0.0000	0.0000	0.0000	1.0000
0.0000	0.0000	0.0000	1.0000
0.0000	0.0000	0.0000	1.0000
0.0000	0.0000	0.0000	1.0000
0.0000	0.0000	0.0000	1.0000
0.0000	0.0000	0.0000	1.0000
0.0000	0.0000	0.0000	1.0000

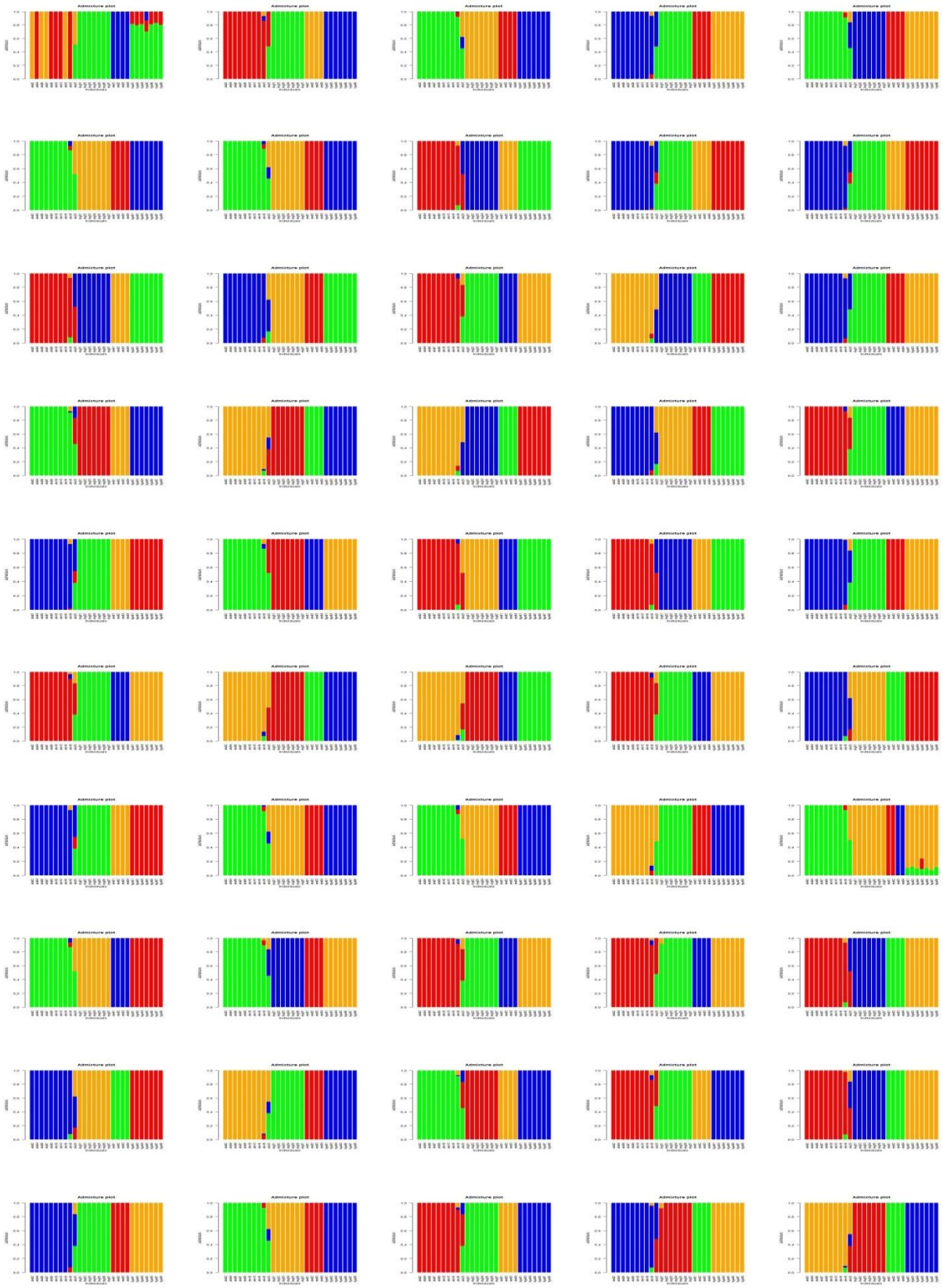
Supplementary figure 2 – 100 plots of all the bootstrap replicates from the two admixture analyses. All vertical lines along the x-axis indicate the individuals. The y-axis indicates the admixture proportions. Four populations are defined by each color, but the order of the populations is random.

a) The first 50 bootstrap plots includes the three side-striped jackals.

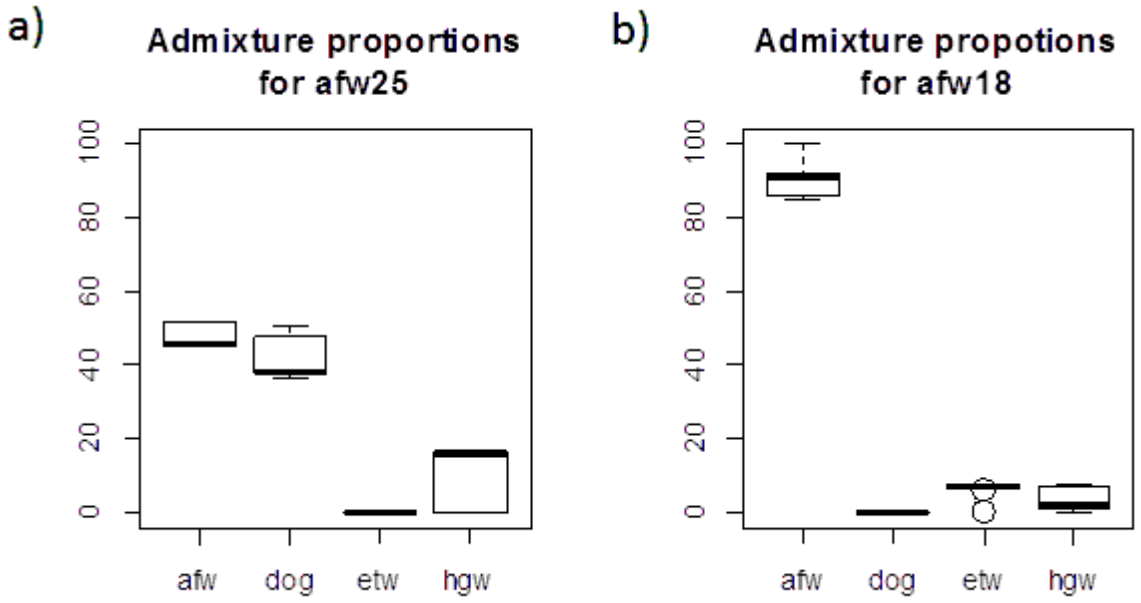


Supplementary figure 2 - continued

b) The last 50 bootstrap plots excludes the three side-striped jackals.



Supplementary figure 3 – Admixture proportions of the two African wolves affected by hybridization or gene flow from sympatric canids based on 50 bootstrap replicates from the admixture run without side-striped jackal.
a) Box plot showing the proportions of "afw25", an African wolf sample collected from Senegal.
b) Box plot shows the proportions of "afw18", an African wolf sample collected from Ethiopia.



Supplementary table 4 – A list of observed F4 values before and after downweighting calculated by the program named F4.

Label	Observed F4	After downweighting	Label	Observed F4	After downweighting
afw02	0.00708	0.00552	**dog05**	0.00694	0.00542
afw04	0.00703	0.00548	**dog06**	0.00690	0.00540
afw06	0.00698	0.00545	**dog07**	0.00667	0.00522
afw07	0.00718	0.00558	**etw01**	0.00677	0.00531
afw09	0.00708	0.00552	**etw02**	0.00674	0.00527
afw10	0.00710	0.00554	**etw03**	0.00703	0.00547
afw13	0.00699	0.00546	**etw04**	0.00697	0.00545
afw16	0.00721	0.00561	**hgw01**	0.00734	0.00571
afw18	0.00690	0.00540	**hgw03**	0.00690	0.00541
afw25	0.00501	0.00403	**hgw04**	0.00707	0.00551
dog01	0.00672	0.00527	**hgw05**	0.00397	0.00324
dog02	0.00662	0.00520	**hgw06**	0.00756	0.00588
dog03	0.00677	0.00529	**hgw07**	0.00701	0.00547
dog04	0.00719	0.00558	**hgw08**	0.00734	0.00572